Development of a Robust Hybrid Estimator using Partial Least Squares Regression and Artificial Neural Networks

Arshad Ahmad[1] **Lim Wan Piang[2]**

[1] *Laboratory of Process Control, Department of Chemical Engineering, Faculty of Chemical and Natural Resources Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia*
Tel: +60-7-553-5610, Fax: +60-7-558-1463, E-mail: arshad@fkkksa.utm.my

[2] *Laboratory of Process Control, Department of Chemical Engineering, Faculty of Chemical and Natural Resources Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia*
Tel: +60-7-553-5858, E-mail: wanpiang@hotmail.com

Abstract

Measurement difficulty is one of the process control issues arising from the complexity and the lack of on-line measurement devices. One of the alternative solutions to deal with this problem is inferential estimation where secondary variables, such as temperature and pressure are used to predict the unmeasured primary variables that are mainly product qualities. This paper presents the estimation of product compositions for a fatty acid fractionation column using a hybrid technique. The proposed technique combines partial least squares regression (PLS) and artificial neural networks (ANN) in an estimation paradigm to provide better estimation properties. The aim is to take advantage of ANN capability to capture the non-linear relationships as well as the statistical strength of PLS method. The results of process estimation using both PLS and hybrid methods are presented. The significant improvement obtained by the hybrid strategy revealed its capability as a potentially viable estimator for product properties in chemical industry.

Keywords:

Inferential Estimation, Partial Least Squares Regression, Artificial Neural Networks, Hybrid Model, Robustness

Introduction

The world trend of chemical production is now moving towards full capacity operation with zero accidents, zero emissions and high profitability. Under this stringent environment, many plants have been forced to revamp their existing control system. However, problems that are related to process dynamic and measurement remain largely unsolved, even when advanced control system is in place. Process dynamic issues are often related to the non-linearity of chemical processes while problems arising from measurement are mainly due to difficulty in implementing on-line measurements. For example, an on-line gas chromatograph, a common instrumentation for the measurement of product compositions, is not suitable for on-line application in many cases. This is due to the low sampling rate and occasional result inconsistency. Furthermore, it is not economic from the viewpoint of operational and maintenance cost. Due to these difficulties, inferential estimation has been recommended as one of the alternative solution.

Inferential estimation is a strategy that employs the measurement of secondary process outputs such as temperatures and pressures, to infer the unmeasurable disturbances on primary process outputs, such as product compositions [1]. For example, intermediate tray temperatures are usually being used to predict product compositions in a distillation column. Hence, the main role of the estimator is to predict the primary variables using selected secondary variables. The estimated values are then fed into the controller for control purposes. For practical implementation, the estimator should provide reliable prediction of the unmeasured.

Background of Inferential Estimation

Since the development of inferential control in 1970s, various approaches to construct the process estimator have been widely studied. The fundamental method is to use the information of the process, such as mass and energy balances, to construct the estimator. Although this is a reliable and direct approach, the developments of such models are laborious and knowledge intensive. For these reasons, researchers have been formulating alternative methods. Most of these methods use the input-output data and some basic knowledge of the process to develop the process estimator, such as Kalman filtering, statistical methods and black box modelling methods.

Application of statistical methods in chemical process modelling and control is not new. These methods include linear regression (LR), multiple linear regressions (MLR), principal component regression (PCR), principal component analysis (PCA) and partial least square regression (PLS). MLR is among the most widely applied methods in chemical industry for estimation. Recently, the PLS method is also

ISBN:983-2643-15-5

gaining popularity in this field compared with more classical MLR and PCR due to its robustness [2]. In process estimation, the use of PLS was pioneered by Mejdell and Skogestad [3,4]. Over the years, they had developed the composition estimator using PCR and PLS models for a binary distillation column. The estimators, which were based on steady-state data and multiple temperature measurements, performed well in various conditions such as multi-component mixtures, pressure variations, and non-linearity. Dealing with the problems of non-linearity and noise in distillation column, they proposed the use of additional factors, weighting functions, and logarithmic transformations.

Some other researchers had also investigated the application of PLS in process estimation and control. Budman and co-workers [5] addressed the development of a robust inferential estimator for a packed-bed reactor by using PLS model. This estimator was then comparing with another estimator developed by using Kalman filter technique. Results showed that PLS estimator was significantly more accurate for estimating the actual concentration in a wide range of operating conditions. Another example is the work by Kresta and co-workers [6]. Their estimator, which was designed to estimate the distillation compositions, had shown good prediction when dealing with large numbers of highly correlated measured variables without over-fitting. They had also proved that the model was more robust to missing data and sensor failures.

The PLS model had been inferior due to its dependency on steady state data and insufficiency when dealing with non-linear system [5]. Efforts to improve this technique in order to deal with both dynamic and non-linear process had been explored. Dayal & MacGregor [7] proposed recursive exponentially weighted PLS algorithm to improve parameter estimation. This newly developed algorithm was tested on a multivariable CSTR and an industrial mineral floating circuit. In the estimation of distillation compositions, Kano and co-workers [8] carried out a comprehensive study of dynamic PLS to improve the accuracy of estimation by using simulated time series data. They concluded that the estimation of top and bottom column quality based on reflux flow rate, reboiler duty, pressure and multiple tray temperatures was much better than the usual tray temperature control system.

Another approach to extend the PLS model in dealing with dynamic and non-linear system is by hybridising the model with other modelling paradigm such as artificial neural networks (ANN). The use of ANN within the PLS modelling paradigm and was first recommended by Qin & McAvoy [9]. The capability of ANN model in dealing with non-linear system had inspired the merging of these methods. Since the results of the NNPLS model were encouraging, some other researchers had also worked on this field to improve the model capabilities. Baffi and co-workers [10] had proposed two extensions models. These were the modified NNPLS and radial basis function network PLS (RBFPLS). Both of these model employed error-based input weights updating procedure to improve the prediction capability. Consequently, Abebiyi & Corripio [11] had also carried out similar investigation to NNPLS model. They proposed a dynamic NNPLS (DNNPLS) in which the static neural network models in the inner relationship were replaced by dynamic neural network models. This approach had been tested with the data from a highly non-linear fluidised catalytic cracking unit and an isothermal reactor. Results showed that the prediction was as good as a MIMO neural network and it was better than PLS-ARMA model.

Partial Least Squares Regression

Partial least squares regression is one of the multivariate analysis methods. It is a linear system identification method that projects the input-output data down into a latent space, extracts a number of principal factors with an orthogonal structure, while capturing most of the variance in the original data [12]. Details description of the PLS structure can be found in [2].

The schematic diagram of the PLS model is illustrated in Figure 1. It consists of two outer relations and an inner relation. The outer relations are the matrixes of independent and dependent variables, which can be represented by **X** and **Y**, respectively. The input **X** is projected into the latent space by the input-loading factor, **P** to obtain the input scores, **T**. Similarly, the output scores, **U** is obtained by projecting the output **Y** into latent space through the output-loading factor, **Q**. These relations are in matrix form and are written in Equation (1) and (2).

Outer relations: $$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E}_f \quad (1)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{F}_f \quad (2)$$

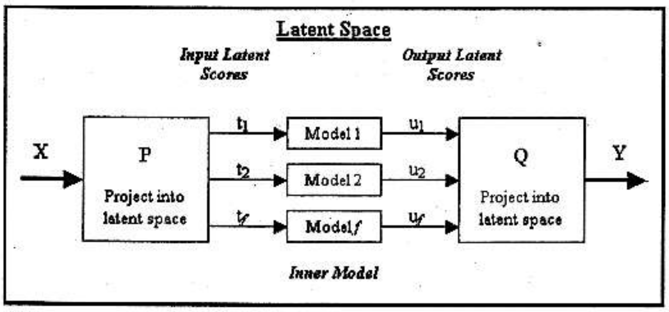

Figure 1 Schematic of the PLS Model [11]

The matrices $\mathbf{E}_f$ and $\mathbf{F}_f$ are residuals of **X** and **Y**, respectively. **X** and **Y** are linked with a linear regression called inner relation to capture the relationship between the inputs and output latent scores. The notation of the inner relation is written in Equation (3).

Inner relation: $$\mathbf{U} = \mathbf{T}\mathbf{B} \quad (3)$$

The procedure of determining the scores and loadings factor is carried out sequentially from the first factor to the *f*th factor. Scores and loading vectors for each factor is calculated from the previous residual matrices as shown in Equation (4) and (5), where initially $\mathbf{E}_0 = \mathbf{X}$ and $\mathbf{F}_0 = \mathbf{Y}$.

ISBN:983-2643-15-5

For **X**, $\mathbf{E}_f = \mathbf{E}_{f-1} - \mathbf{T}_f\mathbf{P}_f^T$ (4)

For **Y**, $\mathbf{F}_f = \mathbf{F}_{f-1} - \mathbf{U}_f\mathbf{Q}_f^T$ (5)

Calculation of the inner and outer relations is performed until the last factor, f or when residual matrices are below certain threshold.

Hybrid PLS-ANN Model

The hybrid PLS-ANN model is constructed based on the NNPLS model [9]. As mentioned before, this technique incorporated feedforward networks into the PLS modelling, where FFN is used to capture the non-linearity in the model while the statistical strength of PLS is maintained.

The schematic diagram of a NNPLS model is depicted in Figure 2. As mentioned above, a conventional PLS model consists of outer and inner relations, where both of these relations are represented in linear form. In NNPLS model, the PLS outer relations are kept linear to transform the original data into score factors (**U** and **T**). On the other hand, neural networks are accomplished in the inner relation as written in Equation (6):

$$\mathbf{u}_f = \mathcal{N}(\mathbf{t}_f) + \mathbf{r}_f \quad (6)$$

where $\mathcal{N}(\bullet)$ stands for the non-linear relation represented by a neural network. Here, the training data is the score factors generated from the outer relations.

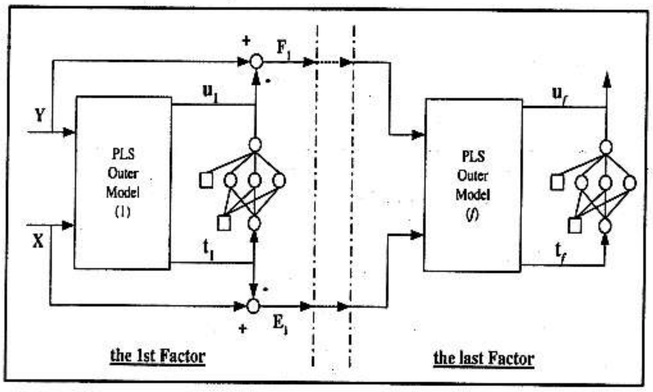

Figure 2 A Schematic Illustration of NNPLS Model [9]

Problem Definition

The aim of this paper is to develop a robust inferential estimator by using hybrid PLS-ANN model based on on-line measurements of process variables, such as flow rates and temperatures. For practical implementation, the estimator should able to provide accurate prediction, and the model must be robust enough to deal with disturbances and changing of operating conditions.

Process Description

The case study considered here is the light cut column of a local fatty acids fractionation plant. At present, indirect control of product compositions is achieved by controlling temperature at selected location. However, this control scheme cannot function very well due to disturbances in the feed composition. This has created some difficulties in the composition control and at times, off-specification products have been produced. In this project, the focus is on the development of a robust inferential estimator for the light cut column.

Light cut column is a packed distillation column consisting of three sections, which are stripping, rectifying and condensing section. This column is operated under vacuum condition induced by steam ejector. The schematic diagram of this column is depicted in Figure 3. The feedstock of light cut column is the bottom product from pre-cut column with fatty acids ranging from C-10 to C-18. The inlet temperature is around 220°C at pressure around 6.84 kPa. Distillate product from this column is C-12 with around 98 %, and the bottom products, which are mainly C-14 to C-18 are then fed to the next column.

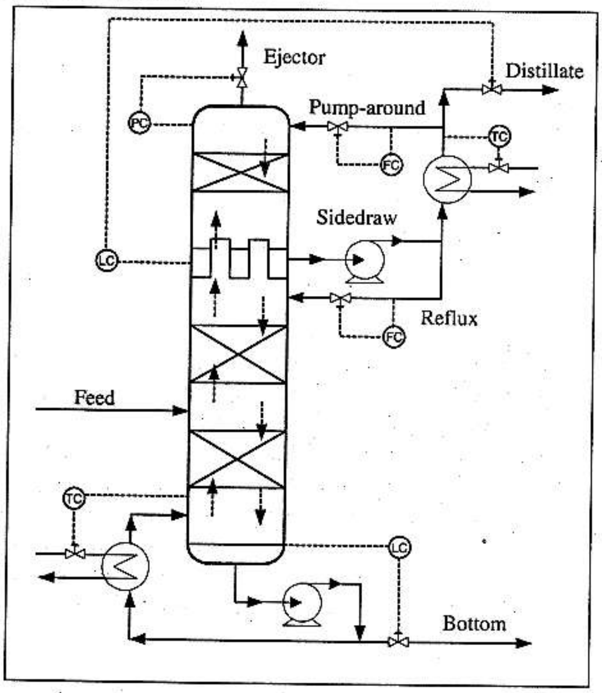

Figure 3 Light Cut Column in the Fatty Acids Fractionation Plant

Dynamic Plant Simulation

The dynamic simulation was carried out using HYSYS.Plant simulator. Based on the process flow diagram provided by a local industry, the light cut column model was set. Here, seven control loops were activated. These are shown in Figure 3. Simulation was carried out in both steady state and dynamic modes. Results of the dynamic simulation were compared to the actual data collected from the plant DCS system. Monitoring and tuning of the control loop was carried out until the simulation results were in close agreement with the actual data.

Sensitivity Analysis

Since the estimation model was data based, selection of

ISBN:983-2643-15-5

appropriate input and output variables is important. Thus, sensitivity analysis of both open and closed loop system was carried out to investigate the dynamic behaviour of process variables such as flow rates, liquid level, temperatures, pressure, and product compositions. This was done by imposing steps changes to various processes input such as temperatures and flow rates. The effects on the process outputs such as tray temperatures and product compositions were then examined. These responses were used as guides to select appropriate input and output variables that are suitable for model development.

An example of the sensitivity analysis results is shown in Figure 4. Here, a 5% increase in the feed temperature was introduced. Results show that the C-12 mole fraction is dropped from about 0.98 to 0.935. It means that the feed temperature has significant effect on the C-12 mole fraction. Based on the results of sensitivity analysis, four input process variables had been selected, namely the feed temperature, the top column temperature, the reflux flow rate and the recycle flow rate. Since tray temperatures had been proven as the secondary variables that are commonly used in inferential estimation [3,8], four tray temperatures had also been chosen. These variables were then used as inputs for the inferential estimator to predict the composition of C-12 fatty acid.

Model Development

In this section, development of the inferential estimator based on both PLS and hybrid PLS-ANN model are described. The performances of these estimators are evaluated on the basis of mean squared error of prediction (MSE) and the explained prediction variance (EPV). The calculations of MSE and EPV are shown in Equation 7 and 8, respectively.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{7}$$

$$EPV = \left\{1 - \frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right\} \times 100\% \tag{8}$$

Here, x is the measurement of the product composition, $\hat{x}$ is its estimation value, $\bar{x}$ is the mean value of measurements, and N is the number of measurement.

PLS Estimator

The strength of PLS model is its capability to deal with a large set of correlated data. For the unity of the data, the selected input variables should be mean-centred and variance scaled through Equation 9 and 10, respectively.

$$z_m = z - \bar{z} \tag{9}$$

$$z_v = \frac{z_m}{\left[\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2\right]^{1/2}} \tag{10}$$

Here, z is an input variable, $\bar{z}$ is the mean value of the input set, z_m is a mean centred value, z_v is the mean centred and variance scaled value, and N is the number of inputs in a data set.

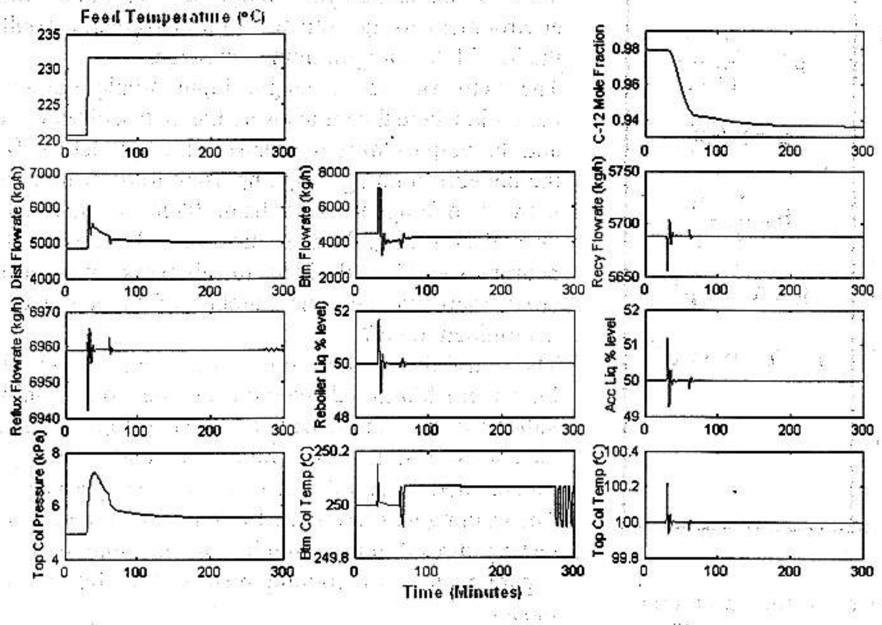

Figure 4 Dynamic Response of 5% Increase in the Feed Temperature

ISBN:983-2643-15-5

The development of the inferential model was carried in the MATLAB environment. The NIPALS algorithm of PLS, which is shown in Table 1, was transferred to the platform of MATLAB using its programming language. The model was first trained using a training data in order to obtain the associate score factors. The numbers of latent variable were set at 20. After the training, the score factors were kept, and they were further used to cross validate different sets of operating data. These results are shown in Figure 5.

Table 1 NIPALS Algorithm of the PLS Model [2,10]

Step	Summary of Steps	
0	Mean centre and scale X and Y	
1	Set the output scores u equal to a column of Y	
2	Compute input weights w by regressing X on u	$\mathbf{w}^T = \dfrac{\mathbf{u}^T \cdot \mathbf{X}}{\mathbf{u}^T \cdot \mathbf{u}}$
3	Normalise w to unit length	$\mathbf{w} = \mathbf{w}/\lVert\mathbf{w}\rVert$
4	Calculate the input scores t	$\mathbf{t} = \dfrac{\mathbf{X} \cdot \mathbf{w}}{\mathbf{w}^T \cdot \mathbf{w}}$
5	Compute output loadings q by regressing Y on t	$\mathbf{q}^T = \dfrac{\mathbf{t}^T \cdot \mathbf{Y}}{\mathbf{t}^T \cdot \mathbf{t}}$
6	Normalise q to unit length	$\mathbf{q} = \mathbf{q}/\lVert\mathbf{q}\rVert$
7	Calculate new output scores u	$\mathbf{u} = \dfrac{\mathbf{Y} \cdot \mathbf{q}}{\mathbf{q}^T \cdot \mathbf{q}}$
8	Check convergence on u. If yes go to step 9 else go to 2	
9	Calculate the input loadings p by regressing X on t	$\mathbf{p}^T = \dfrac{\mathbf{t}^T \cdot \mathbf{X}}{\mathbf{t}^T \cdot \mathbf{t}}$
10	Normalise p to unit length	$\mathbf{p} = \mathbf{p}/\lVert\mathbf{p}\rVert$
11	Compute inner model regression co-efficient b	$\mathbf{b} = \dfrac{\mathbf{t}^T \cdot \mathbf{u}}{\mathbf{t}^T \cdot \mathbf{t}}$
12	Calculate input residual matrix	$\mathbf{E} = \mathbf{X} - \mathbf{t} \cdot \mathbf{p}^T$
13	Calculate input residual matrix	$\mathbf{F} = \mathbf{Y} - \mathbf{b} \cdot \mathbf{t} \times \mathbf{q}^T$
14	If additional PLS dimensions are necessary, replace X and Y by E and F, respectively and repeat steps 1 to 13	

In order to evaluate the performance of the inferential estimator, the model was tested on three sets of data. They were made up of different operating conditions:

- Data A – Normal operating conditions
- Data B – Intermediate fluctuations
- Data C – Severe fluctuations

The purpose of the evaluation was to investigate the accuracy and robustness of the model. The actual values and the prediction results of these data are plotted in Figure 6, 7 and 8.

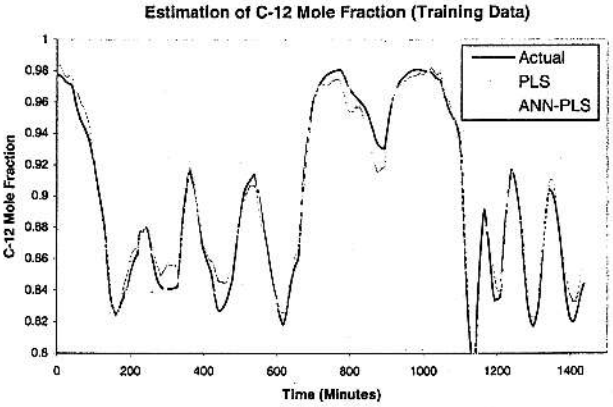

Figure 5 Training Results by Using PLS and Hybrid PLS-ANN Models

Formulation of Hybrid PLS-ANN Model

As mentioned earlier, the PLS model is a linear identification method. In order to improve the ability of the model to deal with non-linear system, a hybrid model, called hybrid PLS-ANN model were formulated. A feed forward network with one hidden layer was incorporated into the PLS model. Hence, it replaced the linear inner model and includes the non-linear feature in the PLS model. Similar to the PLS model, the hybrid model was built in the MATLAB environment using both the Neural Network Toolbox and the MATLAB programming language.

The network was a single input single output (SISO) network, where the inputs were the matrix of score factors, **T**, and the outputs were the matrix of score factors, **U**. Before the network training, it is important to determine the 'best' network topology to avoid the problems of either over-fitting or under- fitting. Hence, the optimal number of hidden neurons should be decided. In this work, we used trial and error approach, and the number of hidden neurons was determined to be 7.

The training algorithm of this network was Levenberg-Marquardt method. For network training, cross validation was implemented as the stopping criteria. The data set was split into a training set and a testing set. The trained model was validated with the testing set sequentially. The training was terminated when the prediction error of the testing dipped into a minimum and started to increase. Figure 5 shows the training results of the hybrid PLS-ANN model.

Similar with the PLS estimator, the hybrid estimator was tested on three sets of data, which were Data A, B and C to evaluate its performance. The predicted C-12 compositions

ISBN:983-2643-15-5

of these data are also plotted in Figure 6, 7 and 8. Subsequently, the performance of both PLS and hybrid model was compared with the actual values.

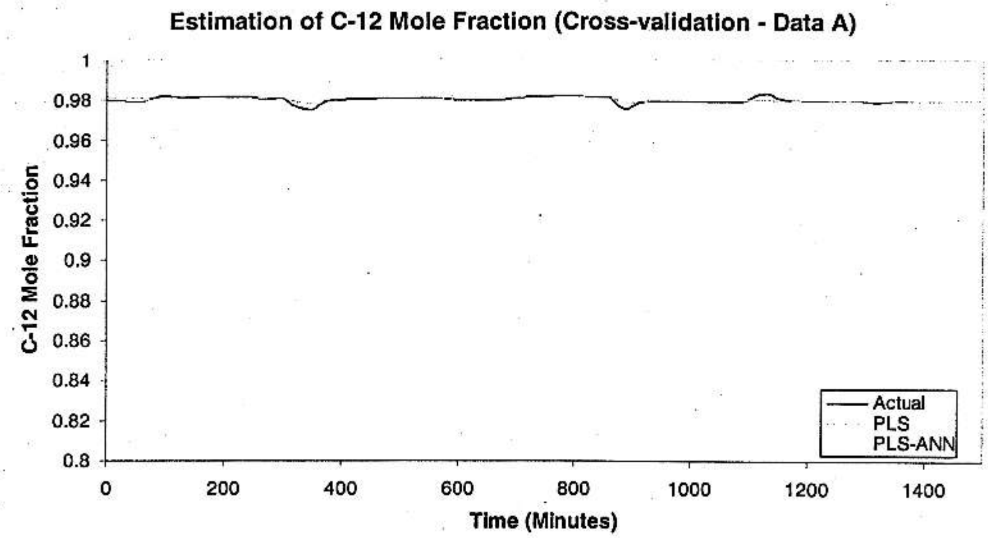

Figure 6 Estimation Results of Data A by Using PLS and Hybrid PLS-ANN Models

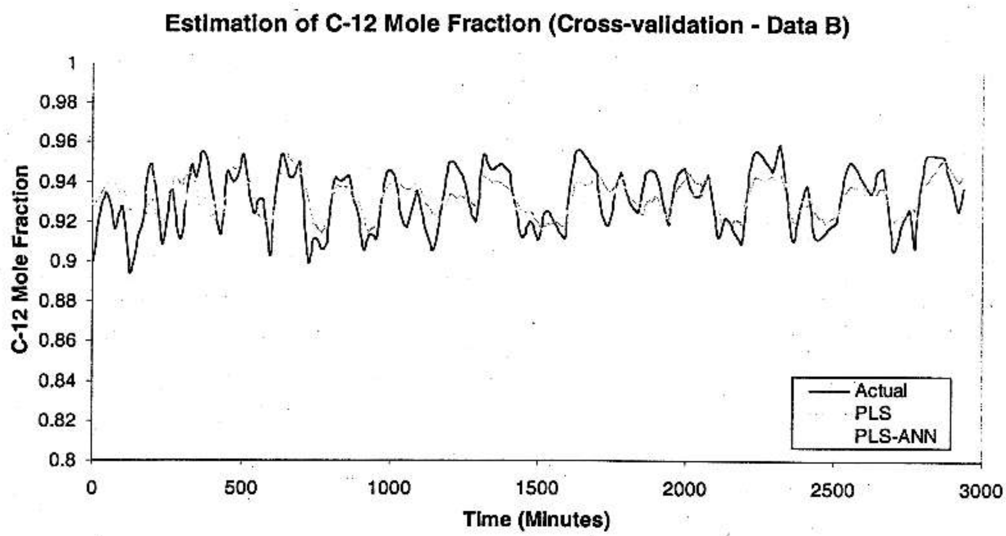

Figure 7 Estimation Results of Data B by Using PLS and Hybrid PLS-ANN Models

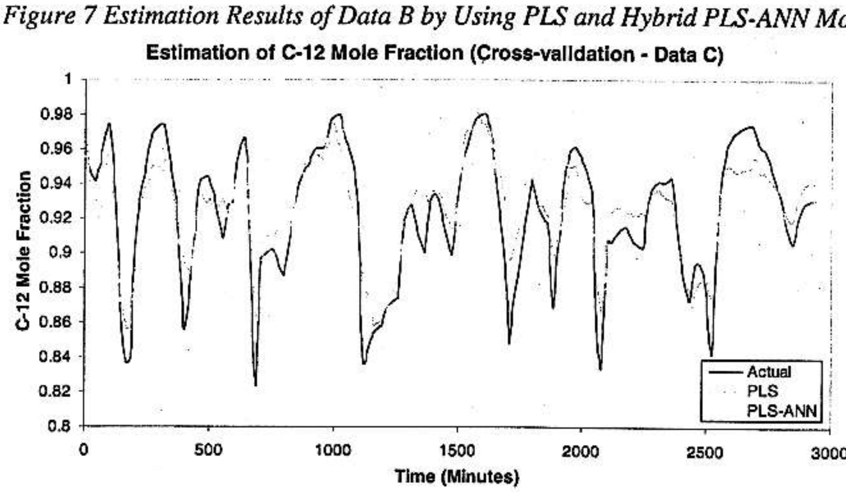

Figure 8 Estimation Results of Data C by Using PLS and Hybrid PLS-ANN Models

ISBN:983-2643-15-5

Discussions

The mean square errors of the training and validation data for both PLS and hybrid estimator were summarised in Table 2. For the training data, the optimum number of latent variables that can best train the data was 20. However, only 5 latent variables were needed to train the hybrid model. It was because the MSE for the cross validation test was not decreased and the EPV for the output data was not remarkable increased using more than 5 latent variables. As a result, the training MSE of PLS model was lower with high percentage of EPV.

Table 2 Comparison of MSE and EPV for PLS and PLS-ANN Model.

Data	Evaluation	PLS	PLS-ANN
Training	MSE	6.2275E-05	1.3464E-04
	EPV	99.38%	36.32%
Data A	MSE	7.4642E-07	1.0369E-06
Data A + noise	MSE	3.3148E-05	1.5036E-05
Data B	MSE	1.0836E-04	8.4157E-05
Data C	MSE	2.5631E-04	1.6990E-04

When the PLS model was tested on data with different operating conditions, the prediction results were good and acceptable. Referring to Figure 6, 7 and 8, the predicted compositions were close to the actual value for Data A. For Data B and Data C, the predicted results were not so accurate, but they were still following the trend of the actual values. Thus, we can say that when the fluctuation in the process was increased, the MSE was getting higher as well.

These predicted results can be improved by using the hybrid PLS-ANN estimator. Similar with the PLS model, this model was trained and then tested on various sets of data. Although the training MSE of the hybrid model was higher than the PLS model, it can perform better when cross validation was implemented. Results showed that the cross validation MSE of the hybrid model were lower in all cases. However, this model faced the similar limitations as in the case of the PLS model. The resulting MSE increases with the increase in process fluctuations.

We have also studied the influence of measurement noise to both of these estimators. 10% noise was introduced to three control variables, which were the top column temperature, reflux and recycle flow rate. The MSE of both estimators were still acceptable and the prediction values were following the trend of the actual values. Thus, we can conclude that these estimators have the ability to deal with measurement noise.

Based on the results, we have proven that the PLS inferential estimator was able to give good prediction using the on-line measured process variables. Moreover, the performance can be improved using the hybrid PLS-ANN model. Apart from the statistical methods, artificial neural networks are also the alternative solution to inferential estimator. A conventional three layer feed forward networks can be used to develop the model. It is still able to give proper prediction with acceptable errors. However, when the model is tested with a large set of correlated data, it will provide poor prediction. This is due to the limitation of the network structure, where the data are not auto-correlated. Nevertheless, the limitation can be overcome using different network structure. Recurrent networks, which support the returnable of some data is suspected to give better results.

Since the structure and development of inferential estimators are still immature, this field is still opened for research. Thus, future works can be done using different model structure. Besides that, additional devices, such as filter and bias can be added to the existing model to improve its accuracy and robustness.

Conclusion

In this paper, the inferential estimator for the product composition of a fatty acid fractionation column was built using PLS model. The on-line measured process variables such as tray temperatures, reflux flow rate, recycle flow rate, feed temperature and top column temperature were used to construct the estimator. This estimator had been performing well in various operating conditions. Moreover, it was able to give good prediction under noisy conditions.

The robustness and accuracy of the PLS estimator can be improved by introducing non-linear feature into the model. This paper incorporated ANN into the PLS model to capture the non-linearity that is always exists in chemical processes. The prediction results proved that the performance was better compared with the PLS model.

The hybrid PLS-ANN estimator is therefore concluded to be applicable to chemical processes. However, the understanding of the first principle model and the dynamic behaviour of the process should not be eliminated during the development of inferential estimator. The lacking of the process information may cause to obtain an unreliable estimator.

Acknowledgments

This project is funded by the Ministry of Science, Technology and the Environment through National Science Foundation Scholarships and IRPA research grant. Our heartiest appreciations are for everybody who has directly or indirectly contribute to the success of this project.

ISBN:983-2643-15-5

References

[1] Joseph, B., and Brosilow, C.B. 1978. Inferential Control of Processes. *AIChE Journal* 24(3):485-508.

[2] Geladi, P., and Kowalski, B.R. 1986. Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta* 185:1-17.

[3] Mejdell, T., and Skogested, S. 1991a. Estimation of Distillation Compositions from Multiple Temperature Measurements Using Partial-Least-Squares Regression. *Industrial Engineering Chemical Research* 30: 2543-2555.

[4] Mejdell, T., and Skogested, S. 1991b. Composition Estimator in a Pilot-Plant Distillation Column Using Multiple Temperatures. *Industrial Engineering Chemical Research* 30:2555-2564.

[5] Budman, H.M., Webb, C., Holcomb, T.R., and Morari, M. 1992. Robust Inferential Control for a Packed-Bed Reactor. *Industrial Engineering Chemical Research* 31:1665-1679.

[6] Kresta, J.V., Marlin, T.E., and MacGregor, J.F. 1994. Development of Inferential Process Models Using PLS. *Computers & Chemical Engineering* 18(7):597-611.

[7] Dayal, B.S., and MacGregor, J.F. 1996. Recursive Exponentially Weighted PLS and Its Applications to Adaptive Control and Prediction. *Journal of Process Contro* 7(3):169-179.

[8] Kano, M., Miyazaki, K., Hasebe, S., and Hashimoto, I. 2000. Inferential Control System of Destillation Compositions Using Dynamic Partial Least Squares Regression. *Journal of Process Contro.* 10:157-166.

[9] Qin, S.J., and McAvoy, T.J. 1992. Nonlinear PLS modelling Using Neural Networks. *Computers and Chemical Engineering* 16(4):379-391.

[10] Baffi, G., Martin, E.B., and Morris, A.J. 1999. Non-linear Projection to Latent Structures Revisited (the Network PLS Algorithm). *Computers and Chemical Engineering* 23:1293-1307.

[11] Adebiyi, O.A., and Corripio, A.B. 2003. Dynamic Neural Networks Partial Least Squares (DNNPLS) Identification of Multivariable Processes. *Computers and Chemical Engineering* 27:143-155.

[12] Wold, H. 1985. Partial Least Squares. In Encyclopedia of Statistical Sciences. Vol 6, 584-591. New York: Wiley.