PARTIAL LEAST SQUARES: APPLICATION IN CLASSIFICATION AND MULTIVARIABLE PROCESS DYNAMICS IDENTIFICATION

Seshu K. Damarla
Department of Chemical Engineering
National Institute of Technology, Rourkela, India
E-mail: Seshu.chemical@gmail.com

Naga C. Kavuri
Department of Chemical Engineering
National Institute of Technology, Rourkela, India
E-mail: biochaitanya@gmail.com

K.S. Kaushikaram
Department of Chemical Engineering
National Institute of Technology, Rourkela, India
E-mail: kaushi8128@gmail.com

Madhusree Kundu*
Department of Chemical Engineering
National Institute of Technology, Rourkela, India
*Correspondence Author: Associate Professor. E-mail: mkundu@nitrkl.ac.in
Phone: +91661-2462263, Fax: +91661-2462999

Abstract

Projection to latent structures or partial least squares (PLS) is a multivariable statistical regression method based on projecting/viewing the information in a high-dimensional data space down onto a low dimensional one defined by some latent variables. PLS is successfully applied in diverse fields including process monitoring; identification of process dynamics & control and deals with noisy and highly correlated data, quite often, only with a limited number of observations available. The conventional PLS is suitable for modeling time independent or steady state processes. For modeling dynamic process, the input data matrix (X) is augmented either with large number of lagged input variables (called finite impulse response (FIR) model) or including lagged input and output variables (called auto regressive model with exogenous input, ARX). By combining the PLS with ARX and FIR model structure, non-linear dynamic processes can be modeled. In the present study, PLS algorithm was used for wine classification and identification of the dynamics of a MIMO process.

In the present work, 178 numbers of wine samples possessing 13 number of feature variables were successfully classified using PLS method with minor misclassifications. Before classification the supervised non- hierarchical K-means clustering was used to designate the classes available among the wine samples, hence discrimination. The efficiency of PLS based classifier was compared with those based on unsupervised neural network ART1 (Adaptive Resonance Theory) and supervised neural network PNN (Probabilistic neural network).

In the present work, a non-linear MIMO distillation process (4×4) was identified with reasonable accuracy along with the evaluation of input-output loading matrix which would logically build up the framework for PLS based process controller. The ARX models as well as least squares were used to build up inner relations among the scores. MIMO processes were casted as a series of SISO identification problems.

Key variable: *PLS, MIMO, ARX, FIR, classification, identification, PRBS*

Introduction

Partial least squares is one of the important multivariable statistics to reduce the dimensionality of the plant data, to find the latent variables from the plant data by capturing the largest variance in the data and achieves the maximum correlation between the predictor (X) variables and response (Y) variables. First proposed by Wold [1] PLS has been successfully applied in diverse fields including process monitoring, identification of process dynamics & fault detection and it deals with noisy and highly correlated data, quite often, only with a limited number of observations available. A tutorial description along with some examples on the PLS model was provided by Geladi Kowalski [2]. When dealing with nonlinear systems, the underlying nonlinear relationship between predictor variables (X) and response variables (Y) can be approximated by quadratic PLS (QPLS) or splines. Sometimes it may not function well when the non-linearities cannot be described by quadratic relationship. Qin and McAvoy [3] suggested a new approach to replace the inner model by neural network model followed by the focused R & D activities taken up by several other researchers like Holcomb & Morari; Malthouse et al.; Zhao et al.; Lee et al.) [4-7].

Discrimination is concerned with separating distinct sets of objects (or observations) on a one-time basis in order to investigate observed differences when casual relationships are not well understood. The operational objective of classification is to allocate new objects (observations) to predefined groups based on a few well defined rules evolved from discrimination analysis of such kind of allied group of observations. Neural networks, either supervised or unsupervised have already emerged as an important tool for classification. The wine data set considered resulted into three clusters with *k*-mans clustering. Present work proposed supervised partial least squares (PLS) based classifier, which was dedicated as well; to authenticate specific category of wine samples out of the three catigories of wine sample present. The PLS classifier was compared with PNN andART-1 based classifier.

Kaspar and Ray [8] developed dynamic extension of the PLS models. Kaspar and Ray demonstrated their approach for identification and control problems using linear models. Lakshminarayanan et al. [9] proposed the ARX/Hammerstein model as the modified PLS inner relation and used successfully in identifying dynamic models. For modeling dynamic process, the input data matrix (X) is augmented either with large number of lagged input variables (called finite impulse response (FIR) model) or including lagged input and output variables (called auto regressive with exogenous input, ARX). By combining the PLS with inner ARX/FIR model structure, nonlinear dynamic processes also can be modeled.

In the identification of MIMO processes, a high degree of correlation is often observed between process variables. One way to circumvent the problem is to use the PLS technique. In the present study, PLS algorithm has been used for identification of the dynamics of MIMO process like multivariable complex distillation column ((4×4)).Discrete input output time series data ($X-Y$) were generated by perturbing non-linear process models with pseudo random binary signals. Signal to noise ratio was set to 10 by adding white noise to the data. The ARX model structure implemented with ordinary least squares were used to build up inner relations among the scores of the discrete input-output time series data ($X-Y$). The (4×4) process was identified in latent subspaces with reasonable accuracy.

Partial Least Squares Model

Linear PLS

If two blocks of measurements say $\boldsymbol{X}$ and $\boldsymbol{Y}$ which are highly correlated, it becomes difficult to predict $\boldsymbol{Y}$ space using only the $\boldsymbol{X}$ space and the ordinary least squares technique. $\boldsymbol{X}$ and $\boldsymbol{Y}$ matrices were auto-scaled before projecting them to latent subspaces. PLS model consists of outer relations ($\boldsymbol{X}$ & $\boldsymbol{Y}$data individually) and inner relations that links $\boldsymbol{X}$ data to $\boldsymbol{Y}$data.

The outer relationship for the input matrix and output matrix can be written as

$$X = t_1 p_1 \ + t_2 p_2{}' + \ldots\ldots\ldots\ldots + t_n p_n{}' + E = TP' + E \qquad (1)$$

$$Y = u_1 q_1^{'} + u_2 q_2^{'} + ... + u_n q_n^{'} + F = UQ^{'} + F \qquad (2)$$

Where T and U represents the matrices of scores of X and Y while P and Q represent the loading matrices for X and Y. if all the components are described, the errors E & F become zero. The inner model that relates X to Y is the relation between the scores T and U.

$$\boldsymbol{U = TB} \qquad (3)$$

Where $\boldsymbol{B}$ is the regression matrix. The response $\boldsymbol{Y}$ can now be expressed as:

$$\boldsymbol{Y = TBQ^T + F} \qquad (4)$$

To determine the dominant direction of projection of X and Y data, maximization of covariance within X and Y is used as a criterion.

$$E_1 = X - t_1 p_1^{'} \qquad (5)$$

$$F_1 = Y - u_1 q_1^{'} = Y - t_1 b_1 q_1^{'} \qquad (6)$$

The procedure for determining the scores and loading vectors is continued by using the newly computed residuals till they are small enough or the number of PLS dimensions is required are exceeded. In practice, the number of PLS dimensions is calculated. By percentage of variance explained and cross validation. The irrelevant directions originating from noise and redundancy are left as E and F. The developed PLS model; i.e. equation (4) can be used to predict the response due to some unknown predictor variable.

Dynamic PLS

For incorporation of linear dynamic relationship in a time series data in the PLS framework, the decomposition of X block is given by equation (1), the dynamic analogue of equation (2) is as follows:

$$Y = G_1(t_1)q_1^{T} + G_2(t_2)q_2^{T} + ...G_n(t_n)q_n^{T} + F = Y_1^{\exp} + Y_2^{\exp} + ... + Y_n^{\exp} + F \qquad (7)$$

Where G_i denotes the linear dynamic model identified at each time instant by ARX model as well as FIR model and $G_i(t_i)q_i^T$ is a measure of Y space explained by the i^{th} PLS dimension in latent subspace. G is the diagonal matrix comprising the dynamic elements identified at each of the n^{th} latent subspaces. Fig.1 represents the PLS based dynamics prediction. Equation (8) represents the ARX structure.

$$y(k) + a_1 y(k-1) + a_2 y(k-1) = b_1 x(k-1) + b_2 x(k-2) \qquad (8)$$

Where $y(k)$ =output at k^{th} instant, $x(k)$ =input. The input matrix for ARX based inner models used in this study was

$$X_{ARX} = \{U_{k-1}, U_{k-2}, T_{k-1}, T_{k-2}\} \qquad (9)$$

Finite Impulse Response Model or FIR model was tested for inner model development. The input matrix for FIR models used:

$$X_{FIR} = \{T_{k-1}, T_{k-2}, T_{k-3}, T_{k-4}\} \qquad (10)$$

T and U represents the matrices of scores of X and Y, respectively. The identified process transfer function:

$$G(z) = {}^{U}\!/_{T} \tag{11}$$

The post compensation of U matrix (PLS inner dynamic model output) with loading matrix Q provided the PLS predicted output Y. The input matrix to the PLS inner dynamic model T was generated by post compensating the original X matrix with loading matrix P. Prior to dynamic modeling, order of the model should be selected. It is difficult to choose the order of the model. Autocorrelation signals render a good indication about order that depends on how many past input and past output values taken in the input matrix for FIR and ARX models. The model parameters for both ARX and FIR models were estimated by linear least square technique.

PLS Classifier

178 numbers of wine samples possessing 13 number of feature variables were successfully clustered in to 3 groups by *k*-means clustering. The 3 different classes of water samples were represented as three numbers of X vectors; each of them were having 25 numbers of data for training and 25 numbers for testing containing 13 feature variables. Three numbers of X vectors were regressed by PLS to three numbers of characteristic Y vectors. The regressed, Y vectors then were given a class membership by using three numbers of column vectors ($Y - 1$). The class memberships were encoded in an appropriate indicator matrix with the corresponding minimum element chosen along the column of the concerned ($Y - 1$). The designed PLS classifier was then used for predicting Y s representing unknown sample classes corresponding to test X samples. Fig. 2 shows the classification as well as authentication performance. It has been found that misclassification rate was only 4 %. In the ART-1 network based classifier, 3 numbers of dedicated ART-1 classifiers designed to identify 3 classes of wine samples were designed with 100 % efficiency and performance of one of them is presented in Table 1. The performance of PNN-based classifier is presented in Table 2.

Table 1: Performance of the ART-1 network based classifier

			Test Data set					
			20%	**30%**	**40%**	**50%**	**60%**	**70%**
ART1-1	**Computation Time**	**ρ=0.4**	0.875	1.0046	0.9709	1.0073	0.872	0.9029
		ρ=0.7	0.955	0.9366	1.0522	0.8738	1.0361	0.9216
	Efficiency	**ρ=0.4**	100%	100%	100%	100%	100%	100%
		ρ=0.7	100%	100%	100%	100%	100%	100%

Table 2: Performance of the PNN network based classifier

		Training Set					
		20%	**30%**	**40%**	**50%**	**60%**	**70%**
% Accuracy for test sets	**20%**	--	55.88235	69.69697	66.66667	78.78788	90.90909
	30%	43.75	--	64.58333	70.83333	70.83333	83.33333
	40%	48.78049	56.09756	--	71.95122	79.26829	84.14634
	50%	47.31183	57.6087	70.65217	--	72.82609	83.69565
	60%	46.46465	49.49495	71.71717	69.69697	--	85.85859
	70%	47.72727	54.54545	68.18182	71.21212	74.24242	--

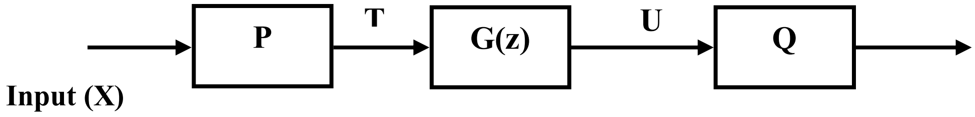

Figure 1 Schematic of PLS based dynamic prediction

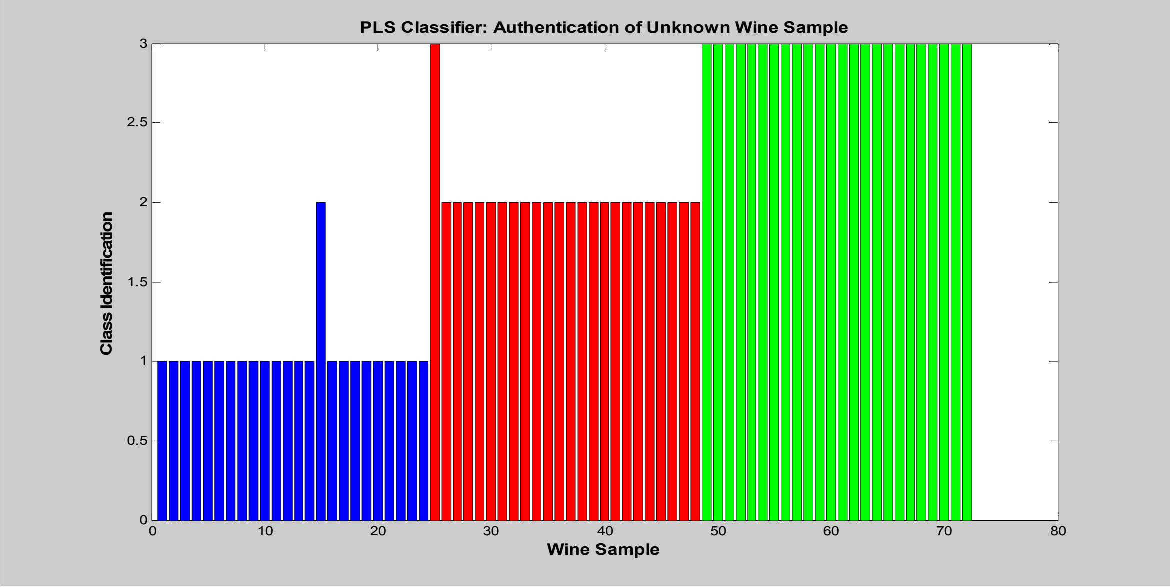

Figure 2: Performance of the PLS classifier developed

Process Dynamics Identification

Complex Distillation Column

Distillation column separates ternary mixture into three products; Top product composition $((XD1))$, Side stream composition $(XS2)$, Bottom product composition $(XB3)$. The four controlled variables including purities of three products and temperature difference between the tray above and below the side tray were controlled by reflux rate, heat input to the reboiler, heat input to the stripper and feed flow rate to the stripper. Using the transfer function (equation (13)), input-output data were obtained by perturbing the process with pseudo random signals (PRBS). Fig.3 demonstrates that identified ARX based PLS predicted dynamics. The FIR based PLS predicted dynamics was comparatively poorer than ARX based PLS predictions. Equation (14) is the representative of the identified ARX based dynamic model.

$$Gp(s)=\begin{bmatrix} \frac{4.09\exp(-1.3s)}{(33s+1)(8.3s+1)} & \frac{-6.36\exp(-1.2s)}{(31.6s+1)(20s+1)} & \frac{-0.25\exp(-1.4s)}{21s+1} & \frac{-0.49\exp(-6s)}{(22s+1)^2} \\ \frac{-4.17\exp(-5s)}{45s+1} & \frac{6.93\exp(-1.02s)}{44.6s+1} & \frac{-0.05\exp(-6s)}{(34.5s+1)^2} & \frac{1.53\exp(-3.8s)}{48s+1} \\ \frac{1.73\exp(-18s)}{(13s+1)^2} & \frac{5.11\exp(-12s)}{(13.3s+1)^2} & \frac{4.61\exp(-1.01s)}{18.5s+1} & \frac{-5.49\exp(-1.5s)}{15s+1} \\ \frac{-11.2\exp(-2.6s)}{(43s+1)(6.5s+1)} & \frac{14(10s+1)\exp(-0.02s)}{(45s+1)(17.4s^2+3s+1)} & \frac{0.1\exp(-0.05s)}{(31.6s+1)(5s+1)} & \frac{4.49\exp(-0.6s)}{(48s+1)(6.3s+1)} \end{bmatrix} \qquad (13)$$

$$G=\begin{bmatrix} \frac{0.01032\,z+0.05828}{z\verb|^|2+1.165\,z-0.1812} & 0 & 0 & 0 \\ 0 & \frac{0.00757\,z+0.06916}{z\verb|^|2+1.432\,z-0.4593} & 0 & 0 \\ 0 & 0 & \frac{0.001917\,z+0.1204}{z\verb|^|2+1.232\,z-0.2472} & 0 \\ 0 & 0 & 0 & \frac{-0.002863\,z-0.005029}{z\verb|^|2+1.083\,z-0.08947} \end{bmatrix} \qquad (14)$$

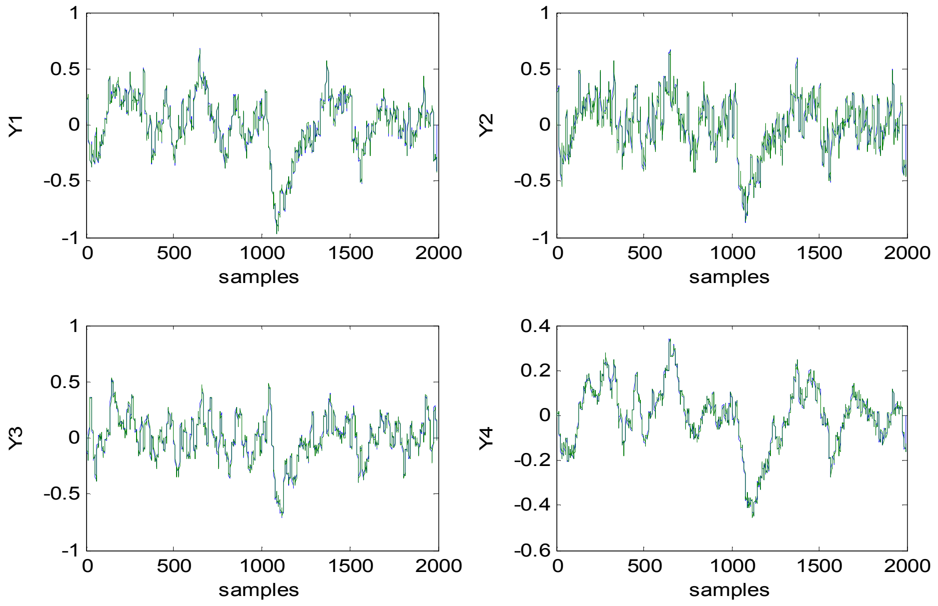

Fig.3: Cross validation: ARX-Model (dashed line) and actual plant (solid line)

Conclusions

The developed PLS classifier was excellent in its authentication performance of unknown wine samples with 4 % misclassifications only. PLS based ARX model could perfectly identified a (4×4) distillation process. The identified latent variable based dynamic model can be used to develop multivariable controllers using loading matrices corresponding to the input and output data matrices.

References

1. H. Wold, *Estimation of principal components and related models by iterative least squares*, In Multivariate Analysis II;Krishnaiah, P. R., Ed.; Academic Press: New York (1966); pp 391-420.
2. P. Geladi, B. R. Kowalski, "Partial least-squares regression: A tutorial," Anal. Chim. Acta, vol. 185, pp. 1-17 (1986).
3. S. J. Qin, T. J. McAvoy, "Nonlinear PLS modeling using neural network," Comput. Chem. Eng., vol. 16 no. 4, pp. 379-391 (1992).
4. T. R. Holcomb, M. Morari, "PLS/neural networks," Comput. Chem.Eng., vol.16 no.4, pp. 393-411 (1992).
5. E. C. Malthouse, A. C. Tamhane, R. S. H. Mah, "Nonlinear partial least squares," Comput. Chem. Eng., vol. 21 no.8, pp. 875-890 (1997).
6. S. J.Zhao, J. Zhang, Y. M. Xu, & Z. H. Xiong, "Nonlinear projection to latent structures method and its applications," Ind.Eng. Chem. Res., vol. 45, pp.3843-3852 (2006).
7. D. S. Lee, M.W. Lee, S. H. Woo, Y. Kim, & J. M. Park, "Nonlinear dynamic partial least squares modeling of a full-scale biological wastewater treatment plant," Process Biochemistry, vol 41, pp. 2050-2057 (2006).
8. M. H. Kaspar, & W. H. Ray, "Dynamic modeling for process control," Chemical Eng. Science, vol. 48 no. 20, pp. 3447-3467 (1993).
9. S. Lakshminarayanan, L. Sirish, & K. Nandakumar, "Modeling and control of multivariable processes: The dynamic projection to latent structures approach," AIChE Journal, vol. 43, pp. 2307- 2323, September (1997).