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.

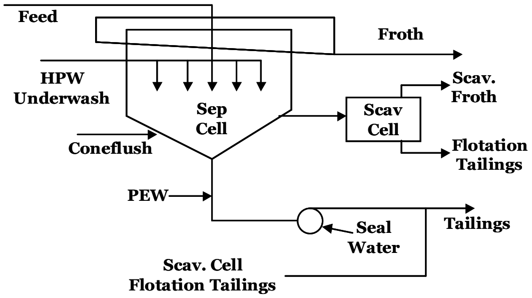

Fig. 1. Process Flow-sheet for Separation Cell

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:

$$Rec = \frac{F_{fr}\rho_{fr}C_{fr}}{F_{fr}\rho_{fr}C_{fr} + F_t\rho_t C_t + F_{ft}\rho_{ft}C_{ft}} \qquad (11)$$

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.

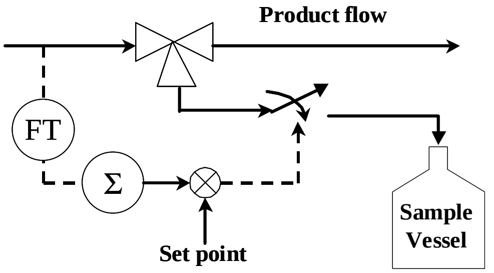

Fig. 2. Sample Consolidation mechanism

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

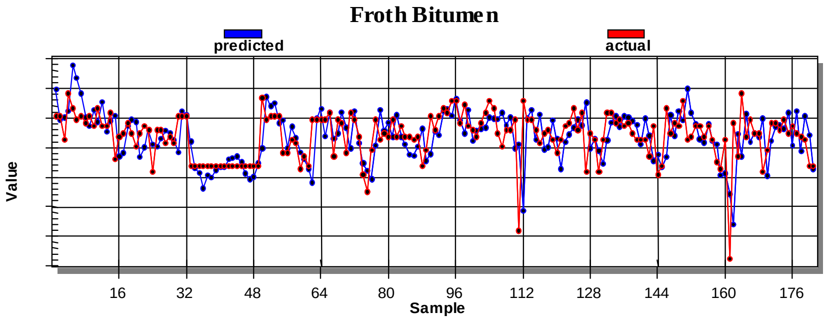

Fig 3. Predictions of Froth Bitumen using PLS Regression

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.

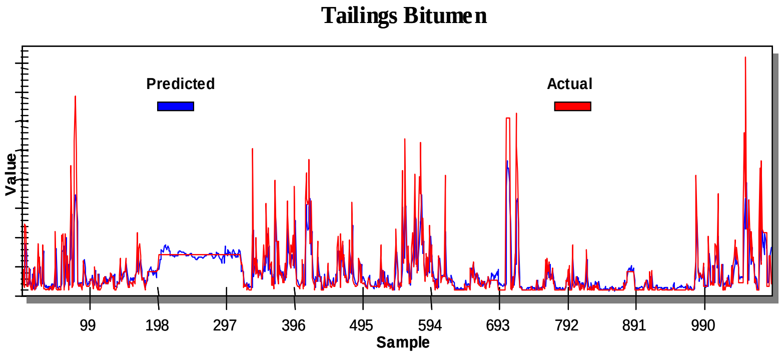

Fig 4. Predictions of Tailings Bitumen using PLS Regression

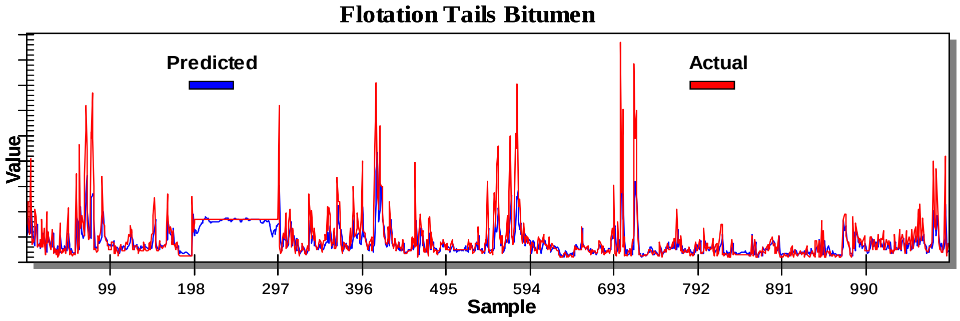

Fig 5. Predictions of Flotation Tails Bitumen using PLS Regression

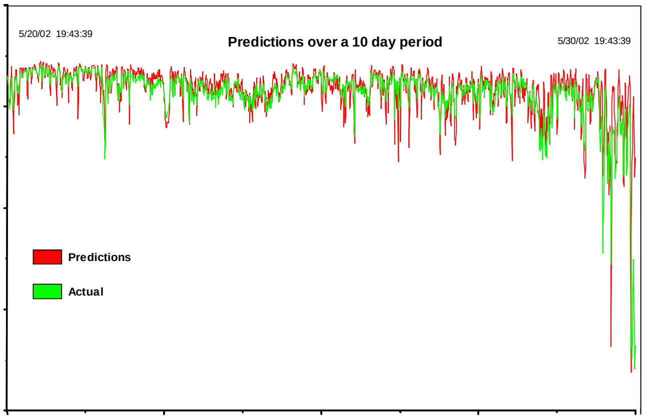

Fig 6. Online prediction of Bitumen Recovery

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.