DATA DRIVEN APPROACH TO VARIABLE SELECTION AND DESIGN OF SOFT SENSORS IN INDUSTRY

Ramesh Kadali [1]

Extraction Process Engineering, Suncor Energy Inc.
P.O. Box 4001, Fort McMurray, AB, T9H 3E3, Canada
rkadali@suncor.com

ABSTRACT

This paper discusses the application of statistical techniques to identify soft sensor models between the process quality variables, whose online measurements are not available in real time, and the process variables measured in real time. The methodology is illustrated through the application of the Canonical Variate Analysis for building soft sensors to predict hydrocarbon compositions in the secondary extraction plant at Suncor.

1. INTRODUCTION

Soft sensors are being increasingly used in process industries where instruments for online measurement of some quality variables are not available. Many numerical methods such as Partial Least Squares, Principal Component Analysis, etc. can be used in the multivariate regression to identify a static model between the quality variables contained in the Y-matrix and the process variables contained in the X-matrix (Raghavan *et al.*, 2002). The real-time predictions of quality variables from these models are called soft sensor predictions and are useful for the control room operators to improve the process performance by taking appropriate corrective actions. Soft sensor predictions can also be used for inferential control (Amirthalingam *et al.*, 2000; Kresta *et al.*, 1994; Li *et al.*, 2002; Parrish and Brosilow, 1985). The notation 'quality' and 'process' variables used in this paper is respectively synonymous with the statistical notation 'response or dependent' and 'predictor or independent' variables.

Modelling soft sensors for industrial application poses some unique data pre-processing challenges such as the presence of missing data and data outliers and may require data filtering. One of the critical steps in building the soft sensor models is the selection of appropriate process variables in the X-matrix. Collinearity of the variables in the X and Y matrices is a problem that needs to be addressed for data regression. Collinearity does not affect the ability of a regression equation to predict the response but gives unreliable estimated of their individual regression coefficients. Hence it poses a problem in estimating the contributions of individual process variables. In this paper discussion on a Canonical Variate Analysis (CVA) based methodology for building soft sensors to provide real time inferential measurements of some quality variables in the oil sands extraction process is provided.

The feed, product and tailings streams from the various unit processes in secondary extraction plant at Suncor Energy have hydrocarbon components A and B, along with water and fine solids. Online measurement of composition using analyzers is a difficult task due to the presence of fine solids. Hydrocarbon composition information in the process streams, which is critical for operating the plant, are currently not available in real time because of the several hours delay caused by sample collection and laboratory analysis. Even though plant personnel need the composition of only one or both of the hydrocarbon components laboratory analysis typically provide composition

[1] The author would like to thank Trevor Hrycay and Ian Noble at Suncor-Extraction for their support.

of all the components in the stream and some additional properties such as the density of hydrocarbons, etc. This additional information is useful in improving the selection of process variables in the X-matrix for building the soft sensors.

The rest of the paper is organized as follows. Section 2 provides the factors that were considered for the selection of the numerical technique for multivariate regression. Section 3 illustrates the basic CVA methodology and section 4 illustrates the industrial application followed by conclusions in 5.

2. SELECTION OF THE NUMERICAL TECHNIQUE FOR MULTIVARIATE REGRESSION

A typical industrial process is multivariate and any quality variable associated to that process is likely dependent on several process variables. Hence building a soft sensor for the quality variable(s) involves identifying a model through multivariate regression. For this we need to explore the relationship between the process matrix $X_{n\times p}$, containing 'n' samples of 'p' process variables, and the quality matrix $Y_{n\times q}$, containing 'n' samples of 'q' quality variables.

At this point we need to choose a suitable numerical technique for data regression to resolve the following problems typically encountered with industrial data.

2.1 Collinearity among the process variable

Since the process variables considered in the X-matrix are from the same plant it is likely that correlations exist between these variables. Collinearity occurs when process variables are so highly correlated that it becomes difficult to distinguish their individual influences on the quality variable(s). Collinearity (ill-conditioning of the X^TX matrix) does not affect the ability of a regression equation to predict the quality variable but it affects the ability to obtain reliable estimates of their individual regression coefficients (Dallal, 2001).

Reduced rank regression techniques can be used to resolve this problem.

2.2 Noise in the variables

Since the measurements of the process variables are coming from instruments they have signal noise. If the measurements are used directly in the model identification step this constitutes an errors in variables (EIV) case. In this paper we consider the case where the process variable measurements are available at a faster time interval t_1 and the quality variable measurements are available at a much slower time interval $t_2 \gg t_1$. The quality variable measurements come from laboratory analysis of the consolidated samples collected over a specific time period through online auto samplers. The process variable measurements can be weighted averaged with respect to the stream flow rate to mimic the auto sampler mechanism. Therefore the j-th sample of the i-th process variable in the X-matrix, represented by x_{ij} where $i = 1, ..., p$ and $j = 1, ..., n$; is obtained by the weighted average

$$x_{ij} = \frac{\sum_{t_2 j}^{k=t_2(j-1)+1} F^k x_{ij}^k}{\sum_{t_2 j}^{k=t_2(j-1)+1} F^k} \tag{1}$$

Since we are using the weighted average data of the process variables for model identification this does not constitute the EIV case.

2.3 Selection of variables in the X matrix

As suggested before, a typical process has multiple process variables and not all of them have significant bearing on the quality variable(s). We want to include only important process variables in the X matrix. This selection can be made based on the relative magnitude of the individual regression coefficients of the process variables.

Consider the case where the composition of only one component in the stream, measured through the laboratory analysis, is of interest to the operators. However laboratory analysis provides data on composition of all the components in the stream and some additional properties. These additional quality variables, although not required for soft sensor modelling in the end, can come in handy for the process variable selection process. We can include all the quality variables obtained from laboratory analysis in the Y-matrix and reduce the number of process variables taken in the X-matrix based on the relative magnitude of the sum of the regression coefficients for all the quality variables.

2.4 Robustness of regression

Since the quality variables from laboratory analysis are available every few hours, building static models to provide soft sensor predictions are considered in this work. CVA methodology, originally developed by (Hotelling, 1935; Hotelling, 1936) to identify associations between two sets of data, appears to be a suitable numerical technique for multivariate regression to obtain steady state models.

3. A PRIMER ON CANONICAL VARIATE ANALYSIS

CVA methodology summarizes the association between two sets of data through a few carefully chosen correlations between linear combinations of variables in the first set and linear combinations of the variables in the second set (Johnson and Wichern, 1982). The pairs of linear combinations are called canonical variables and their correlations are called canonical correlations. CVA methodology first identifies the pair of linear combinations having the largest correlation. Next, it identifies the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair, and the process continues. Interested readers are referred to (Johnson and Wichern, 1982) for detailed derivations.

The linear static model explaining the quality variables Y based on the process variables X is

$$Y = XB + E \tag{2}$$

Where $B_{p\times q}$ contains the model coefficients and E represents the unmeasured disturbance. CVA methodology begins with obtaining the covariance matrix for the joint matrix $\left[\, X_{n\times p} \;\; Y_{n\times q} \,\right]$

$$S = COV(\left[\, X_{n\times p} \;\; Y_{n\times q} \,\right]) = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix} \tag{3}$$

where $S_{xx}(p\times p)$ and $S_{yy}(q\times q)$ contain the covariances between the variables in data sets $X_{n\times p}$ and $Y_{n\times q}$ respectively. $S_{xy}(p\times q) = S_{yx}^T(q\times p)$ contain the covariances between the pairs of variables from different sets $X_{n\times p}$ and $Y_{n\times q}$ and measure the association between the two sets. The linear combinations corresponding to the X and Y matrices are obtained from the eigen vectors of the matrix products $S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx}$ and $S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{yx}$ respectively. The correlations between the pairs of linear combinations are contained in the eigen values. If $Y_{n\times q}$ is the smaller set ($q \leq p$), then q pairs of linear combinations ($U_i(p\times 1)$, $V_i(q\times 1)$ for $i = 1, ..., q$ corresponding to X and Y matrices) are obtained.

Using the first few pairs of linear combinations having significant correlations we can approximate

$$\begin{aligned} \Gamma_{n\times p} &= X_{n\times p}U_1 + ... + X_{n\times p}U_a \\ &= X_{n\times p}U_{p\times a} \end{aligned} \tag{4}$$

$$\begin{aligned} \Upsilon_{n\times q} &= Y_{n\times q}V_1 + ... + Y_{n\times q}V_a \\ &= Y_{n\times q}V_{q\times a} \end{aligned} \tag{5}$$

where $a \leq q$.

The CVA estimate of the model coefficients are obtained as:

$$\hat{B}_{cva} = U(\Gamma^T\Gamma)^{-1}\Gamma^T\Upsilon V^{-1} \tag{6}$$

4. SOFT SENSORS IN SECONDARY EXTRACTION PLANT

Secondary extraction process at Suncor has several unit processes where the feed streams are separated into the hydrocarbon rich product streams, and water and solids in the tailings streams. All the streams contain hydrocarbon components A and B, water and solids to varying degrees of composition. Knowing the composition of each component helps the operators operate the processes more efficiently. Laboratory analysis on 19 of these streams measure several properties such as composition, hydrocarbon density, etc. Operators typically need the information of only one or two of these properties depending on the stream. For example, operators would be interested in knowing the ratio of components A and B in a feed stream, and composition of component B in the tailings stream. However, in soft sensor model building using CVA methodology all the properties measured by the laboratory analysis are included as variables in the Y-matrix to improve the model. Implementing the identified soft sensor models online is in progress.

The different steps of model identification in the case of one of the streams, stream-P16T, are illustrated in figures (1)-(5). In this stream the operators are interested in knowing the composition of components A and B. Figure (1) shows a plot of the data and histograms corresponding to these two variables. From the histograms we can see that the variables are approximately normally distributed. Hence they do not need any non-linear transformation before using the data in model identification.

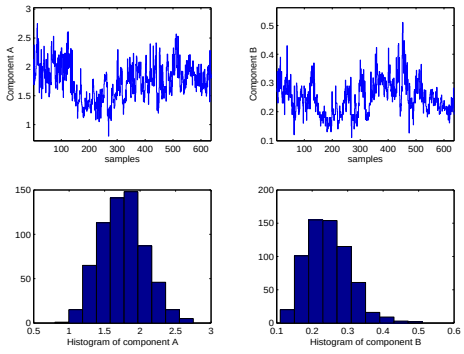

Fig. 1. Data of components A,B in Stream-P16T

A total of 5 laboratory measured quality variables are included in the Y-matrix. Variables Y_1 and Y_2 correspond to compositions A and B in the stream. A total of 16 process variables are included in the X-matrix. Figure(2) illustrates the correlation coefficients corresponding to the

canonical variables between the X and Y matrices. We can see that the first four pairs of canonical variables have significant correlation. Hence these four canonical variables are used in the model identification.

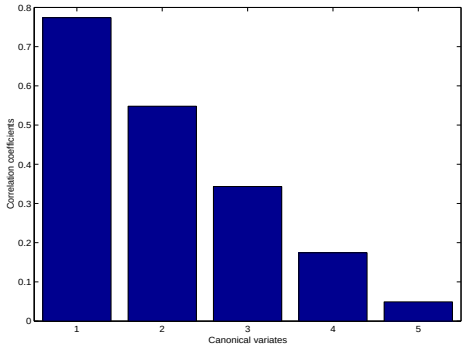

Fig. 2. Correlations corresponding to the canonical variables

The absolute values of the regression coefficients for the variables in X-matrix corresponding to Y_1 and Y_2, and the summation of the absolute values of regression coefficients corresponding to all the 5 variables in Y-matrix are plotted in figure(3). Based on the relative magnitude of the coefficients, 12 variables are selected to be used in the X-matrix for model identification.

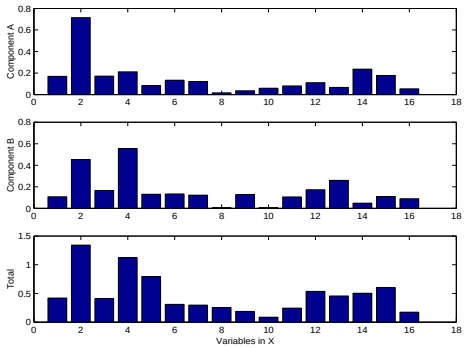

Fig. 3. Absolute values of regression coefficients for variables in X-matrix

A total of 636 samples are available for model building. 486 samples out of these are used for model identification and 150 samples for validation of the models. Figure(4) compares the model predictions with the true values corresponding to Y_1 and Y_2 for the model validation data.

Figure (5) shows the correlation coefficients between the models predictions and true data for the model identification and validation parts of the data for all the 6 variables in Y-matrix. We can see that the model predictions corresponding to Y_1 and Y_2 are showing high correlation with the true data.

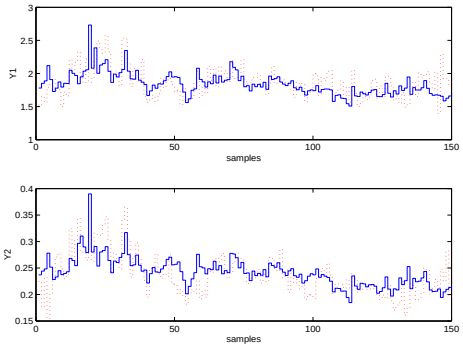

Fig. 4. Comparing the model predictions(-) with true values(..) for validation data

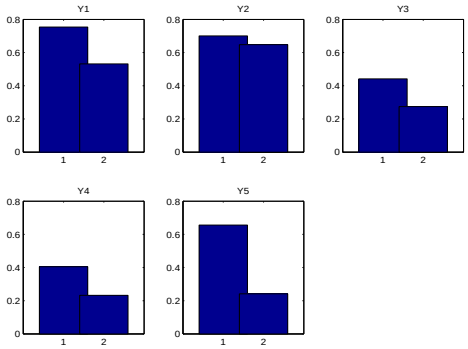

Fig. 5. Correlation coefficients with modelling(1) and validation(2) parts of the data

5. CONCLUSIONS

A CVA based methodology is used to build soft sensors for the estimation of the composition of different components in the secondary extraction process streams. The rationality for the selection of the numerical technique is explained. The methodology is illustrated through the soft sensors design on one of the streams P16T. The soft sensor predictions are found matching well with the laboratory data and online implementation is in progress.

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.

Dallal, G. (2001). The little handbook of statistical practice. In: *http://www.StatisticalPractice.com*.

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology* **26**, 139–142.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika* **28**, 321–377.

Johnson, R.A. and D.W. Wichern (1982). *Applied Multivariate Statistical Analysis*. Prentice-Hall.

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

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

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

Raghavan, H., S. L. Shah, R. Kadali, D. Cox and B. Doucette (2002). Latent variable subspace techniques for the identification of reduced-complexity models using routine operating data. In: *The 52nd CSChE Conference, Vancouver, Canada*.