Proceedings of the 40th IEEE
Conference on Decision and Control
Orlando, Florida USA, December 2001

DEVELOPMENT OF INFERENTIAL DISTILLATION MODELS USING MULTIVARIATE STATISTICAL METHODS

M.A. Evangelista*, F. Neves Jr.†, L.V.R. Arruda♦, A.E.M. Ramos•

CEFET-PR, CPGEI
Av. Sete de Setembro, 3165, 80230-901 Curitiba, PR, BRAZIL.
Tel: +55 41 310-4707 - Fax: +55 41 310-4683
* evange@cpgei.cefetpr.br † neves@cpgei.cefetpr.br ♦ arruda@cpgei.cefetpr.br • ramos@cpgei.cefetpr.br

Keywords: Inferential Models, Partial Least Squares, Multivariate Statistical Methods, Distillation Columns.

Abstract

Chemical processes often have many variables that are being monitored every minute or every second. This can result in "data overload" and useful information that is buried within the collection of data is lost. Techniques that provide a quick method to extract information from large sets of data can prove to be very beneficial. In many cases, however, the data collected from processes are redundant, or highly correlated. In this paper, inferential models for estimating product compositions are built by using Partial Least Squares (PLS) regression, based on simulated time series data. The PLS method removes the correlation problem by projecting the original variable space to an orthogonal latent space. A debutanizer column is used as a case study and the results of the PLS method are compared to another two multivariate statistical methods, which are Multiple Linear Regression (MLR) and Principal Components Regression (PCR).

1 Introduction

For product composition control of distillation columns, it is rarely the case that measurements of product compositions are directly used as controlled variables, because on-line accurate measurements of compositions are difficult (Kano *et al.*, 1998). Most product analyzers, like gas chromatographs, suffer from large measurements delays and high investments and maintenance costs (Mejdell and Skogestad, 1991).

An inferential model is often used in process control when a measurement of the true variable being controlled is not available in real time (Mejdell, 1990). Reasons for the lack of real-time measurements include costs, reliability, and long analysis times or long dead times for sensors located far downstream (Kresta *et al.*, 1994). The general goals for an inferential model are the accurate prediction of the true controlled variable and that these predictions be insensitive for the expected changes around the nominal operating points (Kresta, 1992).

The partial least squares (PLS) regression has been widely applied to chemometrics and chemical industries for data analysis, system identification, and to develop inferential models (Geladi and Kowalski, 1986, Martens and Naes, 1989). Inspired from principal components analysis (PCA) and principal components regression (PCR), the PLS regression is able to give a robust solution in the case of collinear or correlated input variables, where the ordinary least squares regression gives rise to the ill-conditioned problem. The PLS regression removes correlation in the input variable by carrying out orthogonal projections from input space to a latent space. Linear regression is then done in the latent space, which makes the regression solution well defined (Geladi and Kowalski, 1986). According to Kresta, MacGregor and Marlin (1991), the challenges for any multivariate statistical scheme are as follows:

1. The method must be able to deal with collinear data with high dimension, in both process or predictor variables (X) and the product quality or predicted variables (Y).
2. The method must reduce the problem dimension substantially and allow simple graphical interpretation of the results.
3. If both process (X) and product quality (Y) variables are present, it must be able to provide a good prediction of Y.

Principal Components Analysis (PCA) and Partial Least Squares (PLS) are multivariate statistical methods that address the problems mentioned above. PCA is used to explain the variance in a single data matrix by finding combinations of the original variables that represent major trends in the data set. The first principal component (PC) is a linear combination of the columns of X that describes the direction of the greatest degree of variability in the data. The second PC is orthogonal to the first PC and describes the direction of the greatest amount of the remaining variability after the first PC has been calculated. Principal components continue to be calculated until they explain an acceptable amount of variation in the X matrix. With highly correlated data sets, the number of principal components required to explain most of the variation is much less than the number of original variables.

Many statistical methods are available for developing regression models. These methods include Multiple Linear

0-7803-7061-9/01/$10.00 © 2001 IEEE

regression (MLR), Ridge Regression (RR) and Principal Components Regression (PCR). However, the method that appears to best address the problem of high dimensionality and collinearity within the X and Y blocks is Partial Least Squares (PLS). PLS is similar to PCA except that it simultaneously reduces the dimension of the X and Y space to find principal components (often referred to as latent variables in PLS) for the X and Y spaces, which are most highly correlated (Kresta *et al.*, 1991).

PLS regression can be expected to perform better than other regression techniques because of the stability of the predictors or principal components. If the number of variables is too high, the uncertainty of the estimated parameters can become the dominating factor in the variability of the predictors. By using PLS, components are selected that give the maximum reduction in the covariance of the data. In other words, PLS will give the minimum number of variables required to maximize the covariance between the predictor and predicted variables (Höskuldsson, 1988).

The organization of this paper is as follow. Section one discusses the importance of using PLS regression. Section two addresses the problem definition and provides details about the debutanizer process. Section three presents the theoretical foundation of multivariate statistical methods. Section four presents the results with some comments and the final section gives some conclusions.

2 Problem Definition

The problem addressed in this paper is to build inferential models for estimating distillation product compositions based on measurements of process variables, such as tray temperatures, reflux flow rate, condenser pressure, and reboiler heat duty by using multivariate statistical methods. In this section, the example of a debutanizer column and the conditions of steady state and dynamic simulations are illustrated. The inferential models will estimate the mole fraction of pentane (Xc_5) in the distillate stream (butane - Xc_4). The amount of these components is the measure normally used to determine the distillate stream quality.

2.1 Application to a Chemical Process

The steady state and dynamic operations of a debutanizer column were simulated. Hysys.Plant release 2.2 (Hyprotech 2000), which is a rigorous process simulator, was used to simulate the debutanizer process. Hysys.Plant is a fully integrated steady-state/dynamic simulator of chemical and petrochemical processes that allows powerful dynamic modeling capabilities and advanced process control strategies. The schematic diagram of the column is shown in Figure 1. The column consists of 17 theoretical trays including the condenser and the reboiler. The diameter of the column is 2.81 m. The liquid holdups of the condenser and the reboiler are 2.15 m^3 and 3.75 m^3, respectively. The feed stream is a mixture of propanes (small amount), butanes and pentanes and enters the column at the 8th tray. The total flux rate is 8400 Kg/h. Two temperature control loops (TIC-1 and TIC-2), and one condenser pressure control loop (PIC-1) were used in the dynamic simulations. The temperatures on the 4th and 12th trays and condenser vessel pressure were used as controlled variables. Reboiler heat duty, condenser duty, and reflux flow rate were used as manipulated variables. It was used single-loop PID controllers since this is the most common choice in practice, and their parameters were determined by the Ziegler-Nichols method (Ziegler and Nichols 1942).

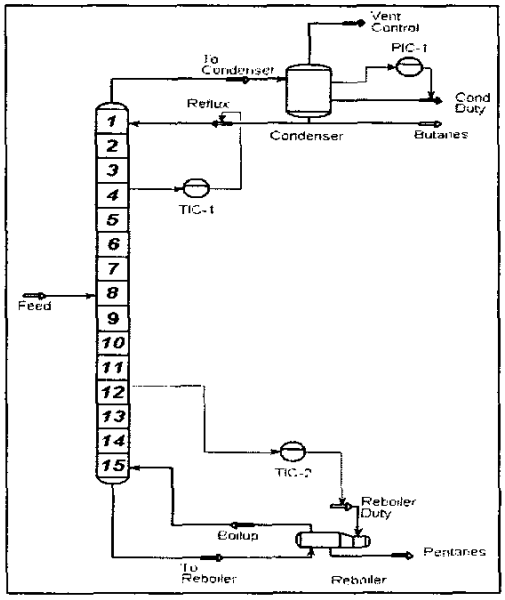

Fig. 1. Schematic diagram of the debutanizer column.

2.2 Conditions of Dynamic Simulations

Appropriate data must be chosen for the reference or calibration set and proper scaling must be used before the principal components analysis can be done. This is critical for the successful application of the procedure (MacGregor et al. 1991). The reference set should consist of information that contains as much of the variance as possible that ultimately leads to successful operation. The conditions of the simulated data used for building the composition inferential models are summarized in Table 1. The disturbances were varied individually and in combination. The data were obtained through open-loop conditions. In other words, the column is assumed to operate under feedback control. However, the controllers use actual values, that is, there is no feedback from estimated values. In the simulations, all variables were measured every minute. Normal distributed random noise of magnitude 0.2°C was added on all temperatures. The calibration and validation sets were obtained under different magnitudes and combinations.

Table 1: Debutanizer Simulation Conditions

Variable	Steady State Conditions	Disturbance Magnitude
Inputs		
Feed Flowrate	10.000 kg/h	± 12%
Feed Temperature	137.90 °C	± 10%
Feed Composition		
Butanes	0.65	± 10%
Pentanes	0.35	± 10%
Manipulated Variables		
Reboiler Duty	4.886E+06 kJ/h	± 8%
Condenser Duty	3.643E+06 kJ/h	± 6%
Reflux Flowrate	8.400 kg/h	± 6%

3 Multivariate Statistical Methods

3.1 Multiple Linear Regression (MLR)

According to Geladi and Kowalski (1986), the multiple linear regression method can be state as follow. Features are measured for m variables x_j ($j=1$ through m) and for a variable y with the goal to establish a linear or first-order relationship between them. This is represented mathematically as:

$$y = x'b + e \qquad 3.1$$

where x' is a row vector of independent variables, b is a column vector of sensitivities or regression coefficients and e is the scalar residual error.

The preceding equations describe MLR for only one sample and one dependent variable. MLR can be extended to more than one sample as well as more than one dependent variable.

While multiple linear regression is available within many computer software packages and it is fairly simple to calculate, caution must be taken before using it (Walpole, and Myers, 1993). This method often fails because X is ill-conditioned. If there is a great deal of correlation or collinearity within the variables of X, the solution of B will be unstable. In the case of instability, slight changes in X can produce completely different results. Although the calibrations may fit the data very well, they are typically not useful for the prediction of new samples (Wise and Gallagher, 1998). Also this method will fail if there are more variables (columns) than samples (rows). In this case, $(X^TX)^{-1}$ would not exist.

3.2 Principal Components Regression

Principal Components Regression (PCR) is one way to deal with the problem of ill-conditioned matrices (Naes and Martens, 1988; Martens and Naes 1989; Höskuldsson, 1996). Instead of regressing the system properties on the original measured variables, the properties are regressed on the principal component scores of the measured variables, which are orthogonal and, therefore, well conditioned, Thus, X^+ is estimated as:

$$X^+ = P(T^TT)^{-1}T^T \qquad 3.2$$

where T is the matrix of principal component scores and P is the matrix of eigenvectors of the covariance or correlation matrix (Wise and Gallagher, 1998). T and P are determined from principal component analysis decomposition on X.

As in Principal Components Analysis (PCA), the number of principal components to retain in the model must be determined. The purpose of the regression model is to predict the properties of interest for new samples. Thus, it is possible to determine the number of Principal Components (PCs) that optimizes the predictive ability of the model. This is typically done by cross-validation, a procedure where the available data is split between training and test sets. The prediction residual error (PRESS) on the test samples is determined as a function of the number of PCs retained in the regression model formed with the calibration data. The procedure is usually repeated several times, with each sample in the original data set being part of the test set at least once. The total prediction error over the entire test sets, as a function of the number of PCs, is then used to determine the optimum number of PCs, *i.e.* the number of PCs that produces minimum prediction error. If all PCs are retained in the model, the result is identical to that for MLR, at least in the case of more samples than variables. In some sense, it can be seen that the PCR model converges to the MLR model as PCs are added (Wise and Gallagher, 1998).

3.3 Partial Least Squares (PLS)

Partial Least Squares (PLS) regression (Geladi and Kowalski, 1986; Lorber *et al.*, 1987; Martens and Naes, 1989; Höskuldsson, 1996) is related to both PCR and MLR and can be thought of as occupying a middle ground between them. PCR finds factors that capture the greatest amount of variance in the predictor variables (X). MLR seeks to find a single factor that best correlates predictor variables with predicted variables (Y). PLS attempts to find factors, which do both, *i.e.* capture variance and achieve correlation. It is commonly said that PLS attempts to maximize covariance (Wise and Gallagher, 1998). The properly scaled X and Y blocks are decomposed as follows:

$$X = TP^T + E = \sum_{a=1}^{A} t_a P_a^T + E \qquad 3.3$$

$$Y = UQ^T + F = \sum_{a=1}^{Z} u_a q_a^T + F \qquad 3.4$$

One of the common methods of calculating PLS is known as the NIPALS algorithm, which stands for non-linear iterative partial least squares. This algorithm calculates a matrix of scores (T) and a matrix of loadings vectors (P) for the X block in the same manner as PCA and PCR. An additional set of vectors, known as the weights (W), is calculated and used to maintain orthogonality in the scores (Wise and Gallagher, 1998). Scores and loadings, U and Q, respectively, are also calculated for the Y block because the Y block may have more than one variable. The steps carried out in the NIPALS algorithm to obtain the PLS decomposition are the followings (Kresta *et al.*, 1991):

0. Mean center and scale X and Y
1. Set u equal to a column of Y
2. $w^T=u^TX/u^Tu$ (regress columns of X on u)
3. Normalize w to unit length
4. $t=Xw/w^Tw$ (calculate the scores)
5. $q^T=t^TY/t^Tt$ (regress columns of Y on t)
6. Normalize q to unit length
7. $u=Yq/q^Tq$ (calculate the new u vector)
8. Check convergence: if YES to 9, if NO to 2
9. X loadings: $p=X^Tt/t^Tt$
10. Regression: $b=u^Tt/t^Tt$
11. Calculate residual matrices: $E=X-tp^T$ and $F=Y-btq^T$
12. If additional PLS dimensions are necessary then replace X and Y by E and F and repeat steps 1 to 9.

Iterations continue until a stopping criterion is satisfied, or X becomes the zero matrix.

It can be shown that the matrix inverse formed by using PLS can be written as (Wise and Gallagher, 1998):

$$X^+ = W(P^TW)^{-1}(T^TT)^{-1}T^T \qquad 3.5$$

where W, P, and T are calculated as above. It also can be shown that weight vector, w, is the eigenvector of the weighted covariance matrix, X^TYY^TX (Höskuldsson, 1988). The first weight vector, w_1, is the eigenvector of the weighted covariance matrix associated with the largest eigenvalue. The second weight vector, w_2, is the eigenvector of the weighted covariance matrix with the next largest eigenvalue, and so on.

3.4 Model Evaluation

The models can be evaluated by their ability to fit the calibration data and to predict new samples. The root-mean-square error of calibration (RMSEC), which gives a measure of how well the model fits the data, can be calculated as follow:

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad 3.6$$

The root-mean-square error of prediction (RMSEP), which gives a measure of the ability of the model to predict new samples, can be calculated exactly as in Equation 3.6 except that the estimates $\hat{y}_i$ are based on a previously developed model, not one in which the samples to be predicted are included in the model building.

3.5 Pretreatment of Data

Appropriate data must be chosen for the reference or calibration set and proper scaling must be used before the principal components analysis can be done. This is critical for the successful application of the procedure (MacGregor *et. al.*, 1991). The reference set should consist of information that contains as much of the variance as possible that ultimately leads to successful operation. If the reference set does not contain enough variance, the analysis may predict faults too often, and if the reference set contains too much variance, a fault may be overlooked. In industrial applications the reference set will come from past data.

The choice of scaling can also be somewhat cumbersome. The variables that contain the greatest amount of variance will be the most dominant in the first principal component. In other words, variables that have values numerically larger than other variables will have a greater significance in the model. This can be avoided by scaling the data to have a zero mean and unit variance. Other methods of scaling have been studied where physical conditions cause certain variable to be more indicative of quality. If some physical conditions are known to be very important, those variables may be scaled to have a greater variance than the rest of the variables. MacGregor, Kresta, and Marlin (1991) investigated several scaling approaches and made the following observations: "If the inner-relationship between the group of variables must be maintained and they are measured in the same units, then the same scaling factor must be used for the entire group. When the variances are almost the same size, then small adjustments in the scaling have an insignificant effect on the results".

4 Results

The output variable to be estimated was the mole fraction of the heavy key component (pentanes) in the distillate product (Xc_5). For building the inferential models, calibration and validation data were obtained by applying disturbances on the variables summarized in Table 1. The predictor variables used to build the models were the temperatures on the 2nd, 4th, 6th, 8th, 10th, 12th, and 14th trays, also top and bottom temperatures, reflux flow rate, condenser pressure, and reboiler heat duty. The data sets were scaled to have zero mean and unit variance because the variables have different units associated with them. The estimation results during the simulations, for all methods, are shown from Figures 2 to 7.

It can be seen from Figure 8 that the MLR models show the smallest RMSEC, the error of calibration. This indicates that, as expected, the MLR model fits the data best because this method uses all the factors. However, it can be noted from the RMSEP, error of prediction, that the MLR does not predict the best. Here the best models were obtained by PLS and PCR methods. In practice, PCR and PLS generally have similar performance though PLS is usually somewhat better. These results can be based on the fact that the PCR method finds factors that capture the greatest amount of variance in the predictor variables, MLR method seeks to find a single factor that best correlates predictor variables with predicted variables, while PLS method attempts to find factors, which do both, i.e. capture variance and achieve correlation and can be expected to perform better than other regression techniques because of the stability of the predictors or latent variables. The optimum number of principal components (PC) and latent variables (LV) were determined by cross-validation.

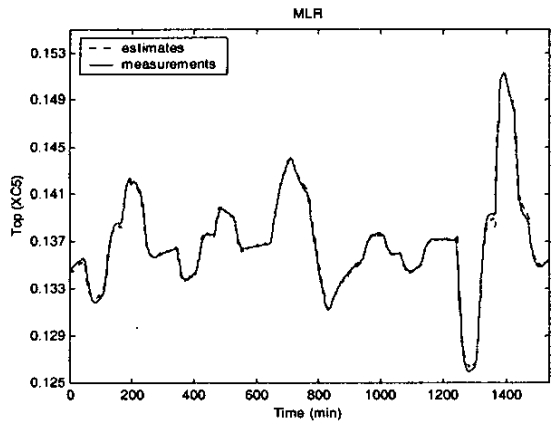

Fig. 2. Estimation results from the calibration set for the pentane molar fraction (Xc_5) in the distillate stream using MLR method.

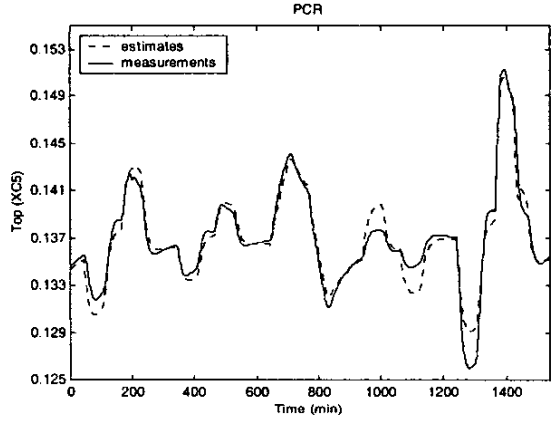

Fig. 3. Estimation results from the calibration set for the pentane molar fraction (Xc_5) in the distillate stream using PCR method.

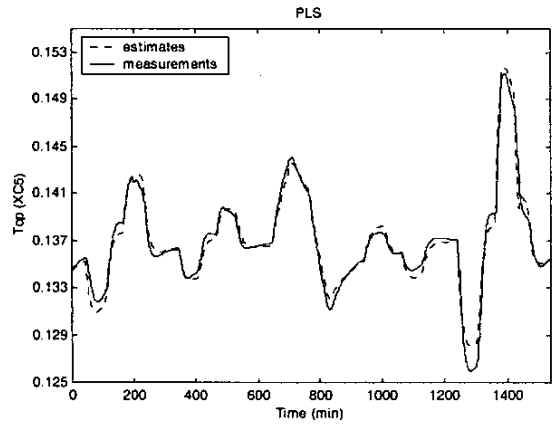

Fig. 4. Estimation results from the calibration set for the pentane molar fraction (Xc_5) in the distillate stream using PLS method.

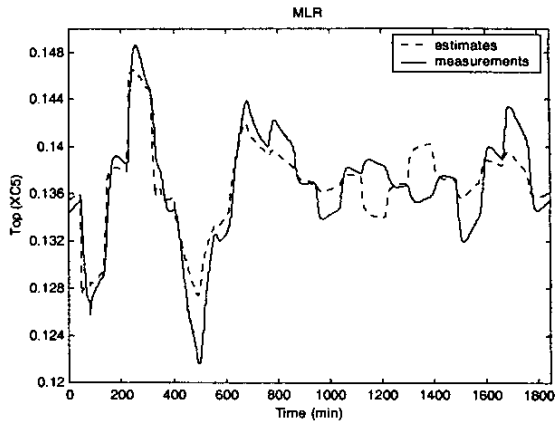

Fig. 5. Estimation results from the validation set for the pentane molar fraction (Xc_5) in the distillate stream using MLR method.

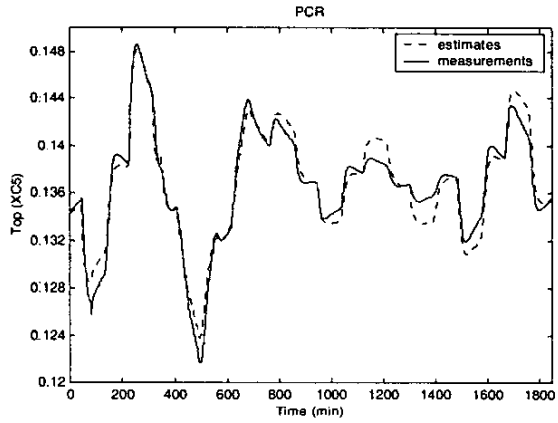

Fig. 6. Estimation results from the validation set for the pentane molar fraction (Xc_5) in the distillate stream using PCR method.

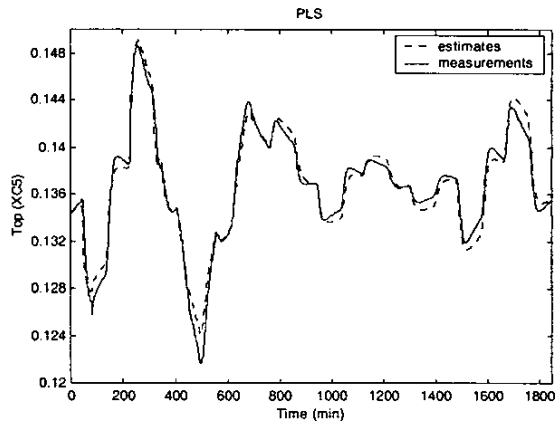

Fig. 7. Estimation results from the validation set for the pentane molar fraction (Xc_5) in the distillate stream using PLS method.

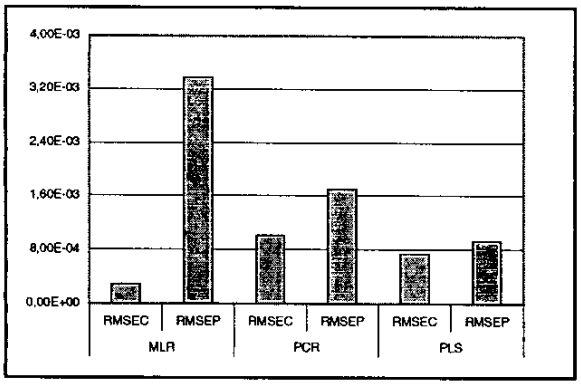

Fig. 8. Calibration (RMSEC) and validation (RMSEP) errors for all methods.

5 Conclusions

In the present work, inferential models, which can estimate the product compositions of a debutanizer column was built by using multivariate statistical methods, such as multiple linear regression (MLR), principal components regression (PCR), and partial least squares regression (PLS). If a system to be identified has correlated inputs, ordinary least squares methods are ill defined and cannot give a robust process model. PLS has shown to be a very powerful approach to building such models when there are large numbers of highly correlated measured variables. The reason why the PLS estimator performed so well, in the debutanizer case, is that there is a very close relationship between the predictor variables and composition. By retaining all measurements without overfitting the data, PLS is able to use all relevant information from the calibration sets, and gives very good predictions on the validation sets. Future refinements of the inferential models should include on-line updating and robust checking of the data and the estimator performance before using them in inferential control loops.

Acknowledgements

We are grateful for the financial support from the Brazilian National Agency of Petroleum (ANP) through its petroleum and gas human resource development program (PRH-ANP-FINEP/MME/MCT 10A CEFET-PR).

References

Geladi, P. & Kowalski, B. (1986). Partial Least-Squares Regression: A Tutorial. *Analyt. Chim. Acta*, 185, 1-17.

Höskuldsson, A. (1988). PLS Regression Methods. *Journal of Chemometrics*, Vol. 2, 211-228.

Höskuldsson, A. (1996). *Prediction Methods in Science and Technology*. Thor Publishing, denmark.

Hyprotech (2000). *Hysys.Plant release 2.2*, Calgary, Canada.

Kano, M., Miyazaki, K., Hasebe, S. & Hashimoto, I. (1998). Inferetnial Control System of Distillation Compositions Using Dynamic Partial Least Squares Regression. *Proceedings of 5th IFAC Symposium on Dynamics and Control of Process Systems (DYCOPS-5)* Corfu, Greece, June 8-10, 377-386.

Kresta, J.V., Marlin, T.E. & MacGregor, J.F. (1991). A General Method for the Development of Inferential Control Schemes Using PLS. *Preprints of 4th International Symposium on Process Systems Engineering (PSE)*, Vol. II, 14.1-14.14, Quebec, Canada.

Kresta, J.V. (1992). The Application of Partial Least Squares to Problems in Chemical Engineering. Ph.D. Thesis, Dept. of Chemical Engng., MacMaster University, Canada.

Kresta, J.V., Marlin, T.E. & MacGregor, J.F. (1994). Development of Inferential Models Using PLS. *Computers Chem. Engng.*, Vol. 18, No. 7, 597-611.

Lorber, A., Wangen, L.E. & Kowalski, B.R. (1987). A Theoretical Foundation for the PLS Algorithm. *J. Chemometrics*, 1 (19).

MacGregor, J.F., Marlin, T.E., Kresta, J. & Skagerberg, B. (1991). Multivariate Statistical Methods in Process Analysis and Control. *Proc. of the Chemical Process Control Conference CPC-IV*, S. Padre Island, Feb. 18-22.

Martens, H. & Naes, T. (1989). *Multivariate Calibration*. Wiley, New York.

Mejdell, T. (1990). Estimators for Product Composition in Distillation Columns. Ph.D. Thesis, The Norwegian Institute of Technology, University of Trondheim, Trondheim.

Mejdell, T. & Skogestad, S. (1991). Estimation of Distillation Compositions from Multiple Temperature Measurements Using Partial-Least-Squares Regression. *Ind. Eng. Chem. Res.*, 30, 2543-2555.

Naes, T. & Martens, H. (1988). Principal Component Regression in NIR Analysis: Viewpoints, Background Details and Selection of Components. *J. Chemometrics*, 2.

Walpole, R.E. & Myers, R.H. (1993). *Probability and Statistics for Engineers and Scientists*. (5th ed.) Englewood Cliffs, NJ, Prentice Hall.

Wise, B.M. & Gallagher, N.B. (1998). *PLS Toolbox 2.0 Manual. Manson*, WA, Eigenvector Research, Inc.

Ziegler, J.G. & Nichols, N.B. (1942). Optimum Settings for Automatic Controllers. *ASME Trans.*, 64, 759-768.