A COMPARATIVE STUDY OF SOFT-SENSING METHODS FOR FED-BATCH FERMENTATION PROCESSES

Hongwei Zhang[*], Zoubir Zouaoui[*] and Barry Lennox[#]

[*]*ASC Technology Computing and Science, North East Wales Institute, UK*
[#] *Control System Centre, The University of Manchester, UK*

Abstract: A comparative study of software sensors using Multiway Partial Least Squares and Extended Kalman Filters in an application to a fed-batch yeast fermentation process is presented. The MPLS theory is introduced firstly and then applied to a yeast fed-batch fermentation process to provide soft-sensing facilities. The soft-sensing capabilities of the MPLS approach are found to compare favourably with the results using EKF. *Copyright © 2005 IFAC*

Keywords: Soft Sensing; Kalman Filters; Statistical Process Control; Fermentation Processes; Batch Control.

1. INTRODUCTION

Industrial fed-batch fermentation systems present a very difficult challenge to control engineers. Problems associated with the nature of the organisms that are being fermented and difficulties related to obtaining accurate information regarding the progression of the batch make controlling and monitoring the process particularly challenging.

The lack of suitable and robust on-line sensors for key fermentation variables such as biomass or product concentration has been considered as a serious obstruction for the implementation of control and optimisation of fed-batch fermentation processes (Aynsley, *et al.*, 1993; James *et al*,. 2002). Considering biomass concentration alone, there are typically two methods available to measure this value – direct or indirect methods. To measure the biomass directly, several techniques have been applied: optical density measurements, capacity measurements, high-resolution liquid chromatography (HPLC), nuclear magnetic resonance (NMR), laser cytometry or biosensors (Golobic *et al*,. 2000). In addition to the high costs associated with these measuring devices, their reliability can be poor when applied to large-scale systems (Montague 1997). It is still the case that most industrial fermentation control policies are based upon the use of infrequent off-line assay information for process operator supervision (Dacosta, *et al*,. 1997). The low sampling frequency associated with such measurements and the inevitable delays in taking samples and performing laboratory tests inevitably compromises the quality of control that is possible using such measurements. As a result of this an alternative approach, that of indirect measurement has attracted a great deal of attention over the last 20 years or so. Indirect measurements of biomass are mathematical algorithms that can produce estimates of unmeasured biomass concentration using the continuously measured variables such as dissolved oxygen, pH and off-gas concentration. The method of estimating the quality related variables from measurements of secondary variables is referred to as ‘Soft Sensing’ or ‘Inferential Estimation’.

The famous Kalman filter has become a popular approach used for inferential estimation and development of software sensors. The Kalman filter is the optimal state estimator for a linear system when a model for the system together with the knowledge of certain stochastic properties of measurement and disturbance noises is available. The Extended Kalman Filter (EKF) is an adaptation to the nonlinear case of the linear Kalman filter. The EKF method optimally tries to estimate the state of the system by assuming that: (1) the behaviour of the system is described by a non-linear model; and (2) the mean and the covariance of the measurement errors are known. Many successful applications have been reported in the literature, using both simulation and experimental investigations (Leigh and Ng, 1984; Gudi, *et al.*, 1995; Aubrun, et al, 2001).

Multivariate statistical methods based on linear projection, such as Principal Components Analysis (PCA) and Partial Least Squares (PLS), have also attracted considerable interest as a method for producing robust empirical models, particularly when there are high dimensionality and collinearities in the data. PCA and PLS, in particular, and their variations, such as neural network partial least squares (NNPLS) (Qin and McAvoy, 1992) and nonlinear principal components analysis (NLPCA) (Dong and McAvoy, 1994, Park and Han, 2000), have been applied to many practical regression problems to estimate quality related variables in chemical engineering processes such as distillation columns, combustion processes , the paper and pulp making process and polymerisation processes. This paper aims to demonstrate Multiway Partial Least Squares (MPLS)'s ability to develop software sensors for fed-batch fermentation processes in a comparative study with EKF.

2. STATISTICAL MODELLING AND SOFT-SENSING USING MULTIWAY PARTIAL LEAST SQUARES

2.1 Partial Least Squares

PLS is a system identification tool that is capable of identifying the relationships between cause (X) and effect (Y) variables. The advantage that this approach offers over more traditional identification techniques, such as ordinary least squares, is that it is able to extract robust models even in applications involving large numbers of highly correlated and noisy process variable measurements.

The approach works by selecting factors of cause variables in a sequence that successively maximises the explained covariance between the cause and effect variables. Given a matrix of cause data, **X**, and effect data, **Y**, a factor of the cause data, $\mathbf{t}_k$, and effect data, $\mathbf{u}_k$, is evaluated, such that:

$$\mathbf{X} = \sum_{k=1}^{np<nx} \mathbf{t}_k \mathbf{p}_k^{\ T} + \mathbf{E} \qquad (1)$$

and

$$\mathbf{Y} = \sum_{k=1}^{np<nx} \mathbf{u}_k \mathbf{q}_k^{\ T} + \mathbf{F} \qquad (2)$$

where **E** and **F** are residual matrices, *np* is the number of inner components that are used in the model and *nx* is the number of causal variables. $\mathbf{p}_k$ and $\mathbf{q}_k$ are referred to as loading vectors.

These equations are referred to as the *outer relationships*. The vectors $\mathbf{t}_k$ are mutually orthogonal. These vectors and $\mathbf{u}_k$ are selected so as to maximise the covariance between each pair, ($\mathbf{t}_k$, $\mathbf{u}_k$). Linear regression is performed between the $\mathbf{t}_k$ and the $\mathbf{u}_k$ vectors to produce the inner relationship, such that:

$$\mathbf{u}_k = b_k \mathbf{t}_k + \varepsilon_k \qquad (3)$$

where b_k is a regression coefficient, and ε_k refers to the prediction error. The PLS method provides the potential for a regularised model through selecting an appropriate number of latent variables, $\mathbf{u}_k$ in the model (*np*). The number of latent variables is typically generated through the use of cross validation.

For further details of the PLS algorithm, the reader is referred to Geladi and Kowalski (1986).

2.2 Multiway Partial Least Squares

PLS is a linear tool, which unfortunately limits its effectiveness when applied to non-linear fed-batch processes. Two options exist for improving the capabilities of PLS when applied to fed-batch systems. The first is to develop non-linear counterparts to PLS and the second is to transform the fed-batch data in such a way as to remove the non-linear characteristics (Nomikos and MacGregor, 1994). Although non-linear PLS techniques exist (Qin and McAvoy, 1992), the transformation of batch data has proved to be a more effective option and has been adopted in this investigation. The most common form of data transformation, termed multiway PLS (MPLS), was initially proposed by Nomikos and MacGregor (1994). Since then other researchers have adopted the approach and applied it to a variety of processes. For example, Gallagher *et al*. (1996) applied the technique to monitor nuclear waste storage vessels and Lennox *et al*. (2001) and Lakshminarayanan *et al.* (1996) investigated the detection of faults in fed-batch fermentation processes.

In a fed-batch process, the cause and effect data can be thought of as being in two 3-dimensional arrays of size *nb* × *nx* × *mx* and *nb* × *ny* × *my*, where *nb* is the number of batches for which data is available, *nx* and *ny* are the number of cause and effect variables respectively and *mx* and *my* are the number of observations of the cause and effect variables respectively that are made during a batch. Unfortunately, PLS requires that the cause and effect arrays be two-dimensional. To address this problem the three-dimensional arrays are recast into two-dimensional arrays in a process referred to as *unfolding*. The concept of unfolding is illustrated in figure 1. The original data arrays are unfolded into a cause variable array, **X**, of size $nb \times (nx*mx)$ and an effect variable array of size $nb \times (ny*my)$. It should be noted that the number of observations made of the cause variables need not be equal to the number of observations made for the effect variables. In fact, it is relatively common to have a single effect measurement made during a batch. This measurement is the final product quality taken at the end of the batch. Following the unfolding of the data, it is then possible to apply PLS to the data in the conventional manner.

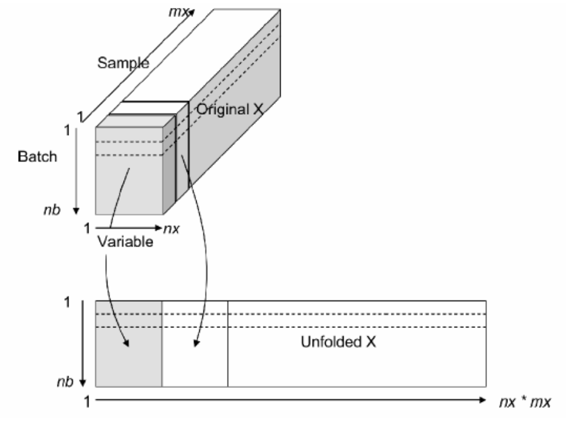


Figure 1 Unfolding

The subsequent use of this model on-line poses the problem that it is necessary to know the values of all process measurements through to the end of the batch, since the unfolded array contains the measurements of each of the variables throughout the duration of the batch. This means that with the exception of the end point of the batch, it is necessary to estimate the future values of all the measured variables. The prediction of future process values is referred to as filling up the array. Of the three methods that were suggested by Nomikos and MacGregor (1994) for filling up the array, Lennox *et al.* (2001) found that the most appropriate method for an industrial fed-batch process was to assume that the values of all process measurements remain at their current offset from the mean trajectory through to the end of the batch. Whilst the most suitable filling up method is likely to be process dependent, this method was also found to be the most appropriate in this work.

The MPLS technique described above and a relatively standard EKF will be applied to a fed-batch yeast fermentation process which is described in the next section.

3. A FED-BATCH YEAST FERMENATION PROCESS

Baker's yeast is a very important micro-organism that has been used for more than a thousand years by human beings. Its importance is illustrated by its use in the baking and brewing industries, in single-cell protein production, and as a host in genetic engineering applications. The yeast used for today's baking, *Saccharomyces cerevisiae*, does not grow during dough raising conditions and therefore it must be supplied from external sources. Since 1917, the fed-batch fermentation technique has been used to produce baker's yeast. The fed-batch culture is an aerobic fermentation process.

The production phase of the yeast production process is mathematically simulated for a fed-batch fermenter with the addition of substrate. The bioreactor was considered completely stirred and isothermal with a variable agitation system and an aeration system. A mechanistic model has been developed which is deterministic and non-structured, based on Monod kinetics with coefficients such as specific rate of growth, specific rate of substrate consumption, and specific rate of oxygen uptake. The model contains 5 mass balance equations and 9 kinetic equations. Initial conditions and all the coefficients are obtained from Bich Pham, *et al.* (1998). Based on the mechanistic model, a simulation has then been built using Matlab to simulate the fed-batch yeast fermentation process. In a typical batch, the biomass concentration increases from 5g/l to about 60g/l in 15 hours. The substrate feed rate follows an exponential curve initially and once it has reached a pre-set maximum value of 0.3 l/h then it is kept constant feed rate to avoid O2 limiting. During the constant feeding phase, the substrate concentration declines and soon falls below the critical value. Then the ethanol peak coincides with the time when the sugar concentration is equal to the critical value. The DOT continues to decrease as long as ethanol is present but rapidly increases when the ethanol is exhausted and stabilises at a concentration corresponding to the feed rate. Simulation results of biomass concentration, substrate concentration, fermenter volume, ethanol concentration, dissolved oxygen tension and substrate flow rate for a typical batch are shown in Figure 2.

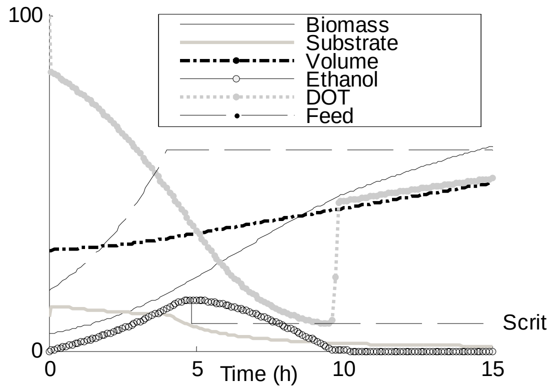


Figure 2 Simulation results of the fed-batch yeast fermentation

4. IMPLEMENTATION OF MPLS AND EKF TO YEAST FERMENTATION

4.1 Implementation of EKF

Since the theoretical properties of EKF are well-understood and can be found in estimation and/or systems theory textbooks (e.g., Brown and Hwang, 1992), it will not be introduced here due to limited space. A relatively standard EKF algorithm is developed and applied to the fed-batch yeast fermentation process to estimate the biomass, substrate, ethanol and DOT concentrations as well as the fermenter volume.

The substrate concentration and the concentration of dissolved oxygen are chosen as observed variables. It is assumed that these variables are measured with a relatively infrequent sampling time at 0.15 h (i.e. 9min) per sample. In experimental works, these measurements are readily available through on-line sensors. The concentration of dissolved oxygen can be measured by a Dissolved Oxygen (DO) probe while a glucose sensor can provide online measurement of substrate (glucose) concentration in the liquid phase. The substrate feed rate is used as an input to the simulation with a sample time of 0.01h. In all the simulations, random deviations of white noise are assumed to exist in the observed variables.

The measurement noise covariance **R** can be obtained from the measurement data and knowledge of the sensor characteristics. The process noise covariance **Q** is usually selected through a trial-and-error procedure using computer simulations. In this work it was shown that, as reported by Oisiovici and Cruz (2000), a well-tuned Kalman filter can be designed by assuming a diagonal and time-invariant process noise matrix. The tuning parameters of the EKF are:

$\mathbf{P_0}$= diag(1, 0.01, 4, 0.1, 10),
Q = diag(10,1,100,100,1);
R = diag(0.01,0.1).

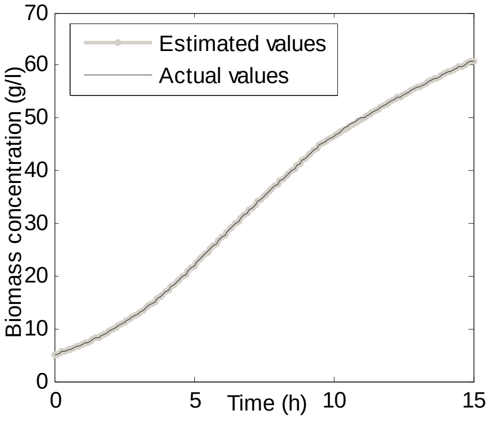
4.2 Implementation of MPLS

To develop a MPLS model based on the simulation, 15 batches of data have been collected. The first 10 of these batches were used to train the model and the remaining 5 were used for validation purposes. When generating the data, the initial conditions and the culture parameters were selected to be the same as those used for EKF. The reason for this is to ensure a fair comparison. Pseuco-Random Binary Signal (PRBS) sequences were applied to the feed rate in order to excite the process so that the data collected had sufficient variation to identify accurate process models.

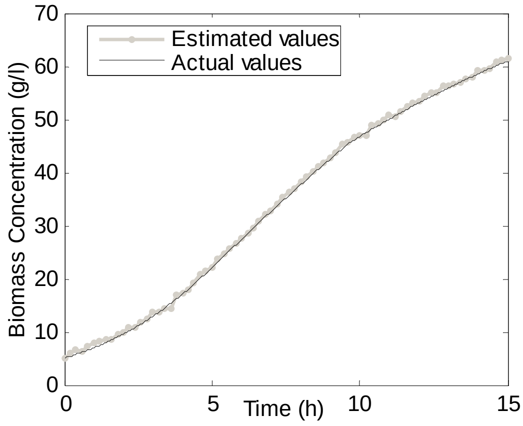
A PLS model, containing 2 latent variables, was identified from the training batches. In this model the following measurements were used as input, or cause, variables: substrate feed rate, dissolved oxygen concentration and culture volume. Based on this model, three software sensors have been constructed to estimate the biomass concentration, substrate concentration and ethanol concentration.

5. COMPARISON OF SOFTWARE SENSORS BETWEEN MPLS AND EKF

The accuracy of the estimates provided by MPLS software sensors are illustrated in figures 3 to 5 which compare the actual concentrations with those predicted by the MPLS model for a typical batch. For comparison purposes, the results on the same process using the EKF approach are also displayed.

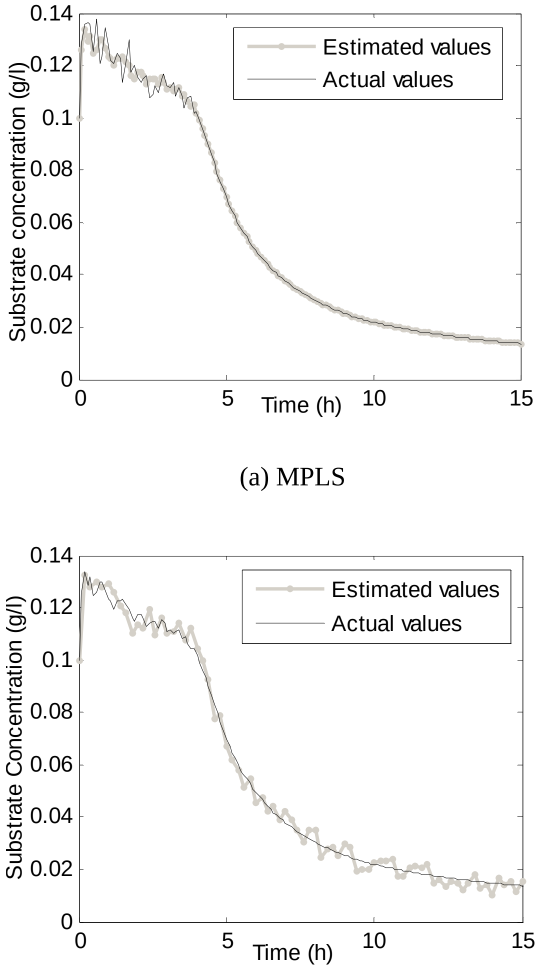


(a) MPLS



(b) EKF

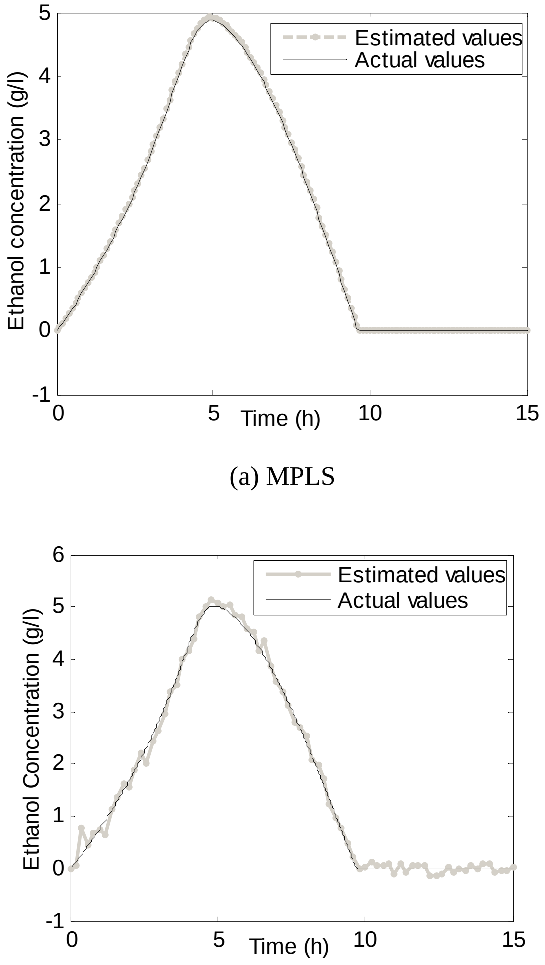
Figure 3 Biomass concentration estimations



(a) MPLS

(b) EKF

Figure 4 Substrate concentration estimations



(a) MPLS

(b) EKF

Figure 5 Ethanol concentration estimations

From the figures above, although both methods provide stable and satisfactory estimations, MPLS provide better estimations than EKF's. This is also demonstrated by the mean square errors (MSE) in figure 6 below. The MSE is calculated using the following equation.

$$\mathrm{MSE} = \frac{\sum_{k=1}^{k_{tend}} \left[x_k - \hat{x}_k \right]^2}{k_{tend}} \qquad (4)$$

where x is the actual value, $\hat{x}$ is the estimated value predicted by either MPLS or EKF and k_{tend} is the total number of sample point in the batch.

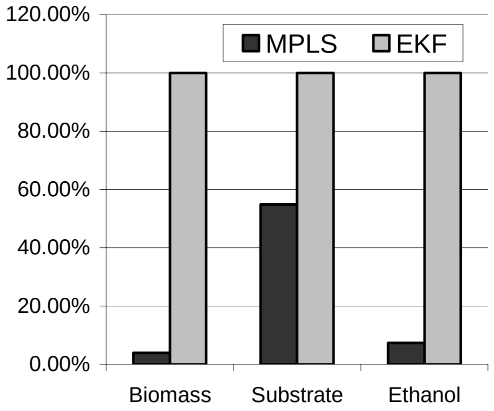


Figure 6 MSE comparison (scaled in percentage)

5. CONCLUSIONS

The results show that MPLS models can be developed for the on-line prediction of low frequency biomass measurements, as well as other variables, using direct secondary measurements. The advantage of the MPLS technique is that it does not need any prior knowledge regarding the process mechanism or the kinetic growth rate, which was required for the EKF model. The disadvantage is that the accuracy of the MPLS model is related to the quality and quantity of experimental data that is available from the process. In many situations this data might be difficult to obtain. In contrast the EKF approach depends largely on the accuracy of the process model. It requires a large design effort and *a priori* estimates of measurement noise and model uncertainty characteristics. The EKF can also suffer from numerical problems and convergence difficulties due to approximations associated with model linearisation. However, it is a generic and elegant approach to cope with the problem of recursive estimation.

REFERENCES

Aynsley M., Holfland A. *et al.* (1993). Artificial Intelligence and the supervison of bioprocesses. *Advances in Biotechnology*. **48**(1): 1-28.

Aubrun C., Theilliol D., Harmand J. and Steyer J. P. (2001). Software sensor design for COD estimation in an anaerobic fluidized bed reactor. *Water Science and Technology*. **43**(7): 115-122.

Bich Pham H. T., Larsson G. and Enfors S.-O. (1998). Growth and energy metabolism in aerobic fed-batch cultures of Saccharomyces cerevisiae: Simulation and model verification. *Biotechnology and Bioengineering.* **60**(4): 474-482.

Brown R.G. and Hwang P.Y.C. (1992). Introduction to random signals and applied Kalman filtering. 2nd edition, *John Wiley & Sons, New York.*

Dacosta P., Kordich C., Williams D. and Gomm J. B. (1997). Estimation of inaccessible fermentation states with variable inoculum sizes. *Artificial Intelligence in Engineering*. **11**(4): 383-392.

Dong D. and McAvoy T. J. (1994). Nonlinear principal component analysis - based on principal curves and neural networks. *Proceedings of the 1994 American Control Conference. Part 2 (of 3), Jun 29-Jul 1 1994*, Baltimore, MD, USA.

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

Gallagher N. B., Wise B. M. and Stewart C.W. (1996). Application of multiway principal components analysis to nuclear waste storage tank monitoring. *Computers and Chemical Engineering*. **20S**: 739-744

Golobic I., Gjerkes H., Bajsic I. and Malensek J. (2000). Software sensor for biomass concentration monitoring during industrial fermentation. *Instrumentation Science and Technology*. **28**(4): 323-334.

Gudi R. D., Shah S. L. and Gray M. R. (1995). Adaptive multirate state and parameter estimation strategies with application to a bioreactor. *AIChE Journal*. **41**(11): 2451-2464.

James S., Legge R. and Budman H. (2002). Comparative study of black-box and hybrid estimation methods in fed-batch fermentation. *Journal of Process Control*. **12**(1): 113-121.

Lakshminarayanan S., Gudi R. D., Shah S. L. and Nandakumar K. (1996). Monitoring batch proceses using multivariate statistical tools: extensions and practical issues. *Proceeding of IFAC World Congress. San Francisco*: 241-246

Leigh J. R. and Ng M. H. (1984). Estimation of biomass and secondary product in batch fermentation. *6th International Conference on Analysis and Optimisation of Systems, Nice, France.*: 19-22

Lennox B., Montague G. A., Hiden H. G., Kornfeld G. and Goulding P. R. (2001). Process monitoring of an industrial fed-batch fermentation. *Biotechnology and Bioengineering.* **74**(2): 125-135.

Montague G. A. (1997). Monitoring and control of fermenters, *Institution of Chemical Engineers*.

Nomikos P. and MacGregor J. F. (1994). Monitoring batch processes using multiway principal component analysis. *AIChE Journal*. **40**(8): 1361-1373.

Oisiovici R. M. and Cruz S. L. (2000). State estimation of batch distillation columns using an extended Kalman filter. *Chemical Engineering Science*. **55**(20): 4667-4680.

Park S. and Han C. (2000). Nonlinear soft sensor based on multivariate smoothing procedure for quality estimation in distillation columns. *7th International Symposium on Process Systems Engineering, Jul 16-Jul 21 2000*. **24**(2): 871-877.

Qin S. J. and McAvoy T. J. (1992). Nonlinear PLS modeling using neural networks. *Computers & Chemical Engineering*. **16**(4): 379-391.