Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Development of Adaptive Soft Sensor Based on Statistical Identification of Key Variables

Ming-Da Ma[1], Jing-Wei Ko[2], San-Jang Wang[3], Ming-Feng Wu[4], Shi-Shang Jang[4]*, Shien-Shu Shieh[5], David Shan-Hill Wong[4]

[1]*Center for Control and Guidance Technology, Harbin Institute of Technology, Harbin 150001, China*
[2]*Refining and Manufacturing Research Center, CPC Petroleum Corporation, Taiwan. Chia-Yi, 600, Taiwan*
[3]*Department of Chemical and Material Engineering, Ta Hwa Institute of Technology, Chiunglin, Hsinchu 307, Taiwan*
[4]*Chemical Engineering Deaprtment, National Tsing-Hua University, Hsin-Chu, 30047, Taiwan*
[5]*Department of Occupational Safety and Hygiene, Chang Jung University, Tainan, 711, Taiwan*

Abstract: An adaptive data-driven soft sensor is derived based on systematic dynamic key variables selection of a process system. The key variables are captured using statistical approaches. The on-line plant measurements can be directly selected as key features to estimate the tardily-detected quality variables. The statistical method adopted is the standard stepwise linear regression. The linear model is adapted as the on-line/off-line quality data becomes available. The adaptation of the model is implemented by standard Kalman filtering theory. The key variables are re-selected in case of new scenarios arrive and are detected by the soft senor. The real time data from an industrial O-xylene purification column is implemented to demonstrate the validity of the approach. Many different scenarios are simulated through an industrial standard dynamic simulator. The simulation results also showed the approach is adequate for the industrial applications.

Keywords: stepwise regression, soft sensor, distillation column, on line adaptation

1. INTRODUCTION

In the last decade real applications for soft sensor has been substantial among the industries (Fortuna, et al., 2007). Many real applications vary from petro-chemical (Badhe, et al., 2007; White, 2003), bio-chemical (Desai et al., 2007), specialty chemicals, (Bhat et al., 2006) to environmental industry (Yoo and Lee, 2004). According to a review by Fortuna et al. (2007), the motivation of the industrial applications of soft sensor has been strengthened because of the need of restriction of product quality and limitation of pollutant emissions.

In the area of soft sensor, the objective is to construct an inferential model to estimate infrequent-measured (lack of on-line sensors) variables using the frequent measured on-line measured data. The advance in the real application is due the fact that more and more real time data can be obtained from the on line sensors. Many researches focused on the data collections and filtering (Lin, et al., 2007), and treatments of data missing (Brás, et al., 2005), data compressions (Nelson, et al., 1996) paved the road of model identification. The soft sensor models implemented can be roughly categorized into the following three types:

(1) Multi-variant statistical models
(2) Artificial intelligence models
(3) First principle models

In case of the last category models, comparatively few works (e.g. Prasad et al., 2002) can be found in the literatures due to fact that these type of models are basically computationally intensive (Lin, et al., 2006). On the other hand, significant parameters are generally unknown (Fortuna, et al., 2007). The great amount of historical data, usually acquired for monitoring purpose, suggests the use of empirical or semi-empirical models. Kano et al. (2000) developed a software sensor based on dynamic partial least squares to predict distillation compositions. Some other works used ARMAX structure nonlinear models and extended them to neural network models to predict distillation column purity. However, as criticized by Kin

(2004), many inferential soft-sensors implemented should be maintained using the laboratory data. Sometimes, even the model structure needs to be checked from time to time. In this work, this problem is taken care based on the combination of adaptive key variables selection and multi-variant statistical model construction.

In the area of multi-variant statistical soft sensor, PCA and PLS are the most useful tools to compress the large amount data into independent information, so that the inferential properties can be predicted from these existed data. Many examples apply PCA/PLS to soft sensors, for instance Zhang et al. (2005) implemented PLS in fermentation process, Skogestad and coworkers (e.g., Mejdell and Skogestad, 1991; Mejdell and Skogestad, 1993) estimated of distillation compositions from multiple temperature measurements. Independent component analysis (ICA) is another relevant technique. A comparison of ICA and PCA is also available by Yoo et al. (2004).

The objective of this work is to derive a novel soft sensor by integrating the statistical key variable selection with an adaptation approach. The superiority of this approach is to narrow down the inferring fast measuring variables to some key variables so that the model becoming easy to maintain. On the other hand, by introducing the concept of adaptation, the model structure will reflect the current scenario of the plant, thus the accuracy of the soft sensor can be substantially improved. This paper is organized as the follows. In section 2, the novel methodology is derived. Section 3 describes the industrial distillation system. The main results are discussed in section 4. Finally, the conclusive remarks are given.

2. THEORETICAL DEVELOPMENT

2.1 Problem Statement

Consider a plant with a set of n online detectable measurement $X=\{x_1(t),x_2(t),\cdots,x_n(t)\}$ and a set of quality variable Y; assume that Y can also be measured with a much slower frequency than X. For the ease of discussion, let's only consider a single quality variable case, $Y = y$. Then, let the plant be expressed by:

$$y(t)=f\left(U(t),D(t)\right) \tag{1}$$

where U are the internal states and D are the slow-varying known/unknown disturbances from outside and t is the time. It is also general that:

$$X(t)=g\left(U(t),D(t)\right) \tag{2}$$

Now, assume that a set of on-line measurement that is collected in the following discrete time form:

$$\Omega=\left\{x_{ij}\,\middle|\,x_{ij}=x_i(t-j\Delta T), i=1,...,n, j=1,...,H\right\} \tag{3}$$

where ΔT is a constant sampling time interval, subscript j indicates the j^{th} observation of the online measurement variable collected in a time horizon H. Let's also assume that a set of quality variable is also available:

$$\Phi=\left\{y_{k+j}\,\middle|\,j=-K_p,...,-1,0,1,...,K_f\right\} \tag{4}$$

where subscript j indicates the j^{th} observation of y. Now, let's separate the observation into the following two categories:

$$\begin{aligned}\Phi_p&=\left\{y\,\middle|\,y_{k-j}, j=1,...,K_p\right\}\\ \Phi_f&=\left\{y\,\middle|\,y_{k+j}, j=1,...,K_f\right\}\end{aligned} \tag{5}$$

where Φ_p indicates the information received before current time, while Φ_f refers the information that will be received in the future. The objective of a soft sensor is to estimate slow measured/unmeasured y using X. Assume that the structure of the soft sensor is in the following form:

$$y_{k+1}=h\left(X_k,X_{k-1},\cdots X_{k-H},\Theta\right)+\Sigma \tag{6}$$

where Θ is a set of parameters, Σ is a set of white noise. The ultimate goal of a soft sensor is to minimize the following:

$$\begin{aligned}&\min_{\Theta}\sum_{j=1}^{K_f}\left(\hat{y}_j-y_j\right)^2\\ &s.t. y_{k+1}=h\left(X_k,X_{k-1},\cdots X_{k-H},\Theta\right)\end{aligned} \tag{7}$$

Since the future information is not available currently, a general approach to develop a soft sensor is based on the current information set Φ_p. The real problem to be solved is as the following:

$$\begin{aligned}&\min_{\Theta}\sum_{j=1}^{K_p}\left(\hat{y}_j-y_j\right)^2\\ &s.t. y_{k+1}=h\left(X_k,X_{k-1},\cdots X_{k-H},\Theta\right)\end{aligned} \tag{8}$$

Let's define that the length of horizon H that the online information is collected be the *regression horizon*. It should be noted that this horizon, as discussed in the next section, should be determined based on the structure of the model. On the other hand, the length of time horizon K_p is the other *identification horizon* to be determined. In this work, a recursive parameter updating algorithm based on Kalman filtering theory is proposed to solve this problem as shown below.

2.2 A Statistical Model Based on Stepwise Regression

In this work, it is assumed that in a particular identification horizon K_p, there exists an optimal regression horizon H such that the following linear model is a solution of (8):

$$\hat{y}(t)=\theta_1 w_1(t-k_1\Delta T)+\theta_2 w_2(t-k_2\Delta T)+\cdots\theta_m w_m(t-k_m\Delta T)+\eta_t \tag{9}$$

where $w_i \in X$, $i=1,...,m$, while η_t should be white noise ideally. In this work, the key variables $w_i \in$W, are determined based on standard statistical stepwise regression as below (Montgomery, 1997):

Step 1: Determine thresholds of probability of type I error (e.g., α_{in}=0.05, α_{out}=0.1), and the corresponding confidence level in F test, (e.g., F_{in}=4.84, F_{out}=3.99).

Step 2: Given a set of $\Psi = \{w_{j1}, w_{j2}, \cdots w_{jk}\}$ are selected in the model (9). If there exists a variable $w_i \notin \Psi$ with a partial F statistics denoted by

$$F_{i|j1,j2,...,jk} = \frac{MSR(i|j1, j2, ..., jk)}{MSE} \qquad (10)$$

is the maximum for all $w_i \notin \Psi$ and

$$F_{i|j1,j2,...,jk} > F_{in} \qquad (11)$$

Then add the w_i into the model (9), and obtain the regression equation and errors using any standard least square regression approach.

Step 3: Consider the new set $\Psi = \{w_{j1}, w_{j2}, \cdots w_{jk}\}$. If there exists a variable $w_i \in \Psi$ with a partial F statistics denoted by

$$F_{i|j1,j2,...,jk-1} = \frac{MSR(i|j1, j2, ..., jk-1)}{MSE} \qquad (12)$$

where $i \neq j1, j2, \cdots, jk-1$. In case,

$$F_{i|j1,j2,...,jk-1} > F_{out} \qquad (13)$$

Then delete the w_i out of the model (9), and obtain the regression equation and errors using any standard least square regression approach.

Step 4: Repeat step 2 and 3 until conditions (10) and (13) are not met for all the variables not selected into the model (9).

2.3Determination of Model Structure

Consider the plant (1), assume that *D(t)* is not changed in a certain period $P > K_p$. Although important key variables can be chosen using the above approach, still some parameters remain underdetermined. Since the model should extracted from the data sets Ω and Φ_p, the size of the data set, i.e. the regression horizon H and identification horizon K_p become critical. It is clear that the process nonlinearity is the major issue of the accuracy of the model. Given Ω and Φ_p, the following optimization problem should be solved:

$$\min_{W,H,K_p} \sum_{i=1}^{K_p} (y_i - \hat{y}_i)^2 \qquad (14)$$

Note that the key variable set is a function of horizons H and K_p, the above problem becomes non-convex. Since the parameter set W can be uniquely determined using the stepwise regression approach in 2-2 for a given horizon H and horizon K_p. Problem (14) can be solved by some heuristic search approach or exhausted search on the feasible plane of H and K_p.

2.4 Online Adaptation of the Model

The process model (9) can be expressed in the following state space form

$$[\alpha_{t+1}] = [T_t][\alpha_t] + [\omega_t] \qquad (15)$$
$$[Y_t] = [Z_t][\alpha_t] + [\upsilon_t] \qquad (16)$$

Here the states represent the relationship between the online measurements X and the slow measured quality variable y. $[\omega_t]$ and $[\upsilon_t]$ are independent, zero-mean, Gaussian noise processes of covariance matrices $[Q]$ and $[R]$, respectively. $[T_t]$, the transition matrix, is a unitary matrix here. $[Z_t]$ is the measurement matrix at time t.

State estimation can then be carried out in a recursive manner from interval to interval. At the start of any time interval t, given an estimated state vector $[\tilde{\alpha}_{t-1}]$ and a estimated covariance matrix of the states $[\tilde{P}_{t-1}]$, then the predicted values of the state vector $[\tilde{\alpha}_{t|t-1}]$ and predicted the covariance matrix for this period $[\tilde{P}_{t|t-1}]$ are given by

$$[\tilde{\alpha}_{t|t-1}] = [T][\tilde{\alpha}_{t-1}] \qquad (17)$$
$$[\tilde{P}_{t|t-1}] = [T][\tilde{P}_{t-1}][T]^T + [Q] \qquad (18)$$

The above equations assume there is no change of the process, and the uncertainty contained in $[P]$ increase.

Once the measurement arrives, the minimum mean square estimator of the states and the covariance matrix can be updated by the following equations

$$[\hat{\alpha}_t] = [\tilde{\alpha}_{t|t-1}] + [\tilde{P}_{t|t-1}][Z_t]^T [\Phi_t]^{-1} \left([\hat{Y}_t] - [Z_t][\tilde{\alpha}_{t|t-1}] \right) \qquad (19)$$
$$[\tilde{P}_t] = [\tilde{P}_{t|t-1}] - [\tilde{P}_{t|t-1}][Z_t]^T [\Phi_t]^{-1} [Z_t][\tilde{P}_{t|t-1}] \qquad (20)$$
$$[\Phi_t] = [R] + [Z_t][P_{t|t-1}][Z_t]^T \qquad (21)$$

These equations ensure that the new estimate of the states reflects the measurement data, and the uncertainty in the covariance matrix decreases.

2.5 The Flow Chart

For the implementation of the proposed adaptive soft sensor algorithm, the flowchart is shown in figure 1. Here, ε is a threshold which determines when the model should be rebuilt. To avoid the affect of measurement noise, it should be bigger than the noise level and can be determined by the operator.

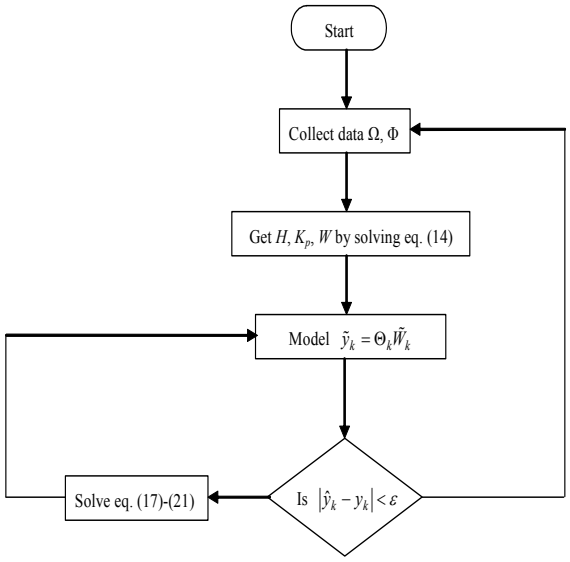


Fig. 1. The flowchart of the proposed algorithm.

3. PROCESS DESCRIPTION

Figure 2 shows a flowsheet of an industrial-scale o-xylene distillation column located in the Refining & Manufacturing Research Institute of CPC Corporation in Taiwan. The 56-stage column, including a total condenser and a partial reboiler, is operated at 1.57 kg/cm^2. Its pressure drop is assumed to be 0.18 kg/cm^2. A fifteen-component mixture with 194.2 kmol/h is fed to the stage 27, numbering from the column top. The column is simulated by using ChemCad. The liquid phase activities were calculated by using SRK. Table 1 shows the fifteen component compositions of feed, distillate, and bottom, respectively at nominal operating condition. Distillate and bottom flow rates are 80.54 kmol/h and 113.66 kmol/h, respectively. It should be noted that the feeding, product, and column specifications are similar to the industrial case. The column is sized as a tray-column with diameter 3 m. The isopropylbenzene (IPB) purity at the distillate of the column is required not to exceed 0.5 mol%. An on-line GC was installed to detect distillate IPB composition. However, composition sensor suffers from measurement delay. In addition, the measurement also suffers from operating perturbations within the column, which result in uncertain indication of the average quality. In order to improve IPB control quality, real-time estimation of IPB purity is required and a composition soft sensor is desired to be developed by temperature measurements. A number of temperature sensors located at stages 1, 13, 23, 31, 44, 45, and 56 are installed in the column to monitor overhead IPB purity.

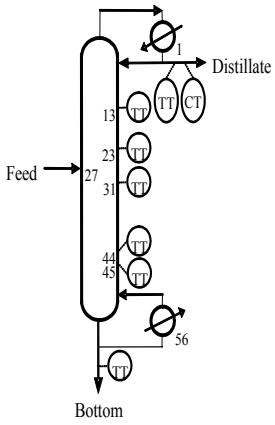


Fig. 2. Flowsheet of an industrial-scale o-xylene distillation column

Table 1 Component compositions at feed, distillate, and bottom

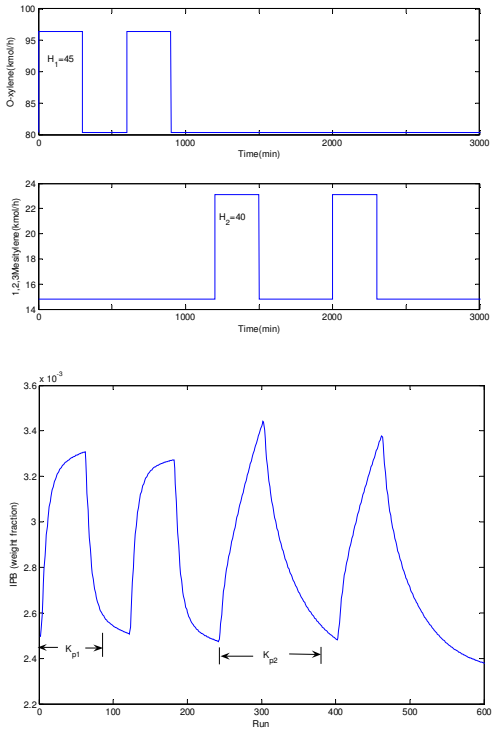
Components	Feed (mol%)	Distillate (mol%)	Bottom (mol%)
n-Nonane	0.07	0.17	9.16e-5
p-Xylene	0.06	0.14	1.02e-5
m-Xylene	2.1	0.51	5.42e-5
o-Xylene	41.34	98.9	0.55
Isopropylbenzene	0.63	0.25	0.90
n-Propylbenzene	1.69	5.87e-3	2.88
1-Methyl-3-ethylbenzene	8.69	9.34e-3	14.84
1-Methyl-4-ethylbenzene	3.92	1.51e-3	6.70
1,3,5-Trimethylbenzene	6.44	5.59e-4	11.00
1,2,4-Trimethylbenzene	21.62	5.62e-5	36.94
1,2,3-Trimethylbenzene	7.61	0.00	13.00
1,2,4,5-Tetramethylbenzene	1.37	0.00	2.34
Naphthalene	0.80	0.00	1.37
1Methylnaphthalene	5.32	0.00	9.09
n-Decane	0.22	1.02e-4	0.38

4. SIMULATION EXAMPLE

Consider the case shown in figure 3, the input components change like figure (3a) and the IPB purity response is shown in figure (3b). Procedures suggested in section 2.5 are implemented and the result is shown in figure 4. It is noted that when the input component 1,2,3-Trimethylbenzene changed at 1200th min, the model structure remained unchanged and large predicting error can be observed. Then

we collect data and rebuild the model, new key variables are selected. The predicting error is small when 1,2,3-Trimethylbenzene changed again at 2000th min.

It can be seen that the proposed algorithm works well when the key variables are selected properly. However, in some cases, when the process has strong nonlinearity and the key variables are not selected correctively, the proposed algorithm may have bad performance or even failed.



(a)

(b)

Fig. 3. The changes of input components and IPB purity.

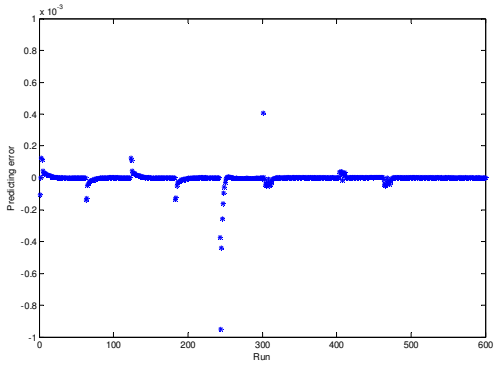


Fig. 4. The predicting error of the IPB purity.

5. INDUSTIAL APPLICATION

It should be noted that distillation is an operation with huge energy consumption. Table 2 shows the energy needed to control the IPB purity at different levels. Therefore, energy consumption can be reduced dramatically if the IPB purity can be estimated precisely and make it remain at higher level while keep it away from warning line in the mean time.

Table 2 Energy consumption with different IPB purity

IPB (%)	Heat duty (kcal/h)	Condenser duty (kcal/h)
0.5	3704000	-5637000
0.4	4200000	-6102000
0.3	4887000	-6820000
0.2	6225000	-8157000
0.1	10260000	-12200000

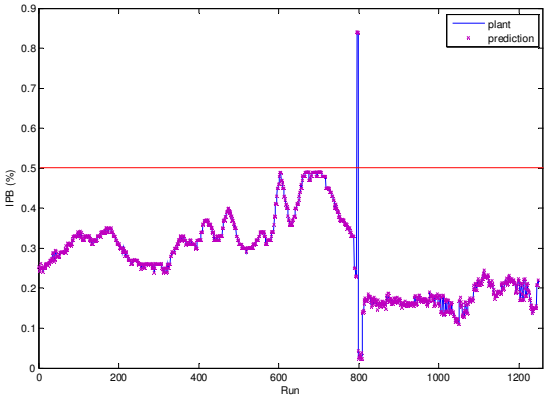


Fig. 5. Industrial application of the proposed method.

In this section, the proposed algorithm is applied to the distillation column of Refining & Manufacturing Research Institute of CPC Corporation in Taiwan. Procedures given in section 2.5 are implemented. The online temperature measurements are selected as key variables and the predicting model based on these key variables is developed. The result is shown in figure 5. It can be seen that the proposed algorithm works well with this example.

6. CONCLUSION

In this paper an adaptive data-driven soft sensor based on statistical identification of key variables is developed. The statistical method adopted is the standard stepwise linear regression. The online plant measurements can be selected as the key variables which make the proposed algorithm easy to maintain. The adaptation of the model is implemented by the Kalman filter. The validity of the proposed method is demonstrated by simulation and industrial examples.

REFERENCES

Badhe, Y.P., J. Lonari, S.S. Tambe and B.D. Kulkarni (2007). Improve polyethylene process control and product quality, Hydrocarbon Processing, March, 53-60.

Bhat S.A., D.N. Saraf, S. Gupta and S.A. Gupta (2006). Use of Agitator Power as a Soft Sensor for Bulk Free-

Radical Polymerization of Methyl Methacrylate in Batch Reactors, Ind. Eng. Chem. Res., 45, 4243 -4255.

Brás L.P., S.A. Bernardino, J.A. Lopes and J.C. Menezes (2005). Multiblock PLS as an approach to compare and combine NIR and MIR spectra in calibrations of soybean flour, Chemom. Intell. Lab Syst., 75, 91-99.

Desai, K., Y. Badhe, S.S. Tambe and B.D. Kulkarni (2005). Soft-sensor development for fed-batch bioreactors using support vector regression. Biochemical Engineering Journal, 27, 225-239.

Fortuna, L., S. Graziani, A. Rizzo, M.G. Xibilia (2007). Soft Sensors for Monitoring and Control of Industrial Processes. Springer, London.

Kano, M., K. Miyazaki, S. Hasebe and I. Hashimoto (2000). Inferential Control System of Distillation Compositions Using Dynamic Partial Least Squares Regression. J. Proc. Cont., 10, 157–166.

Kin, M.J. (2004). How to lose money with inferential properties. Hydrocarbon Processing, October, 47-52.

Lin, B., B. Recke, J.K.H. Knudsen, S. B.Jørgensen (2007). A systematic approach for soft sensor development. Computers and Chemical Engineering, 31, 419-425.

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

Mejdell, T., S. Skogestad (1993). Output estimation using multiple secondary measurements: high-purity distillation, AIChE J., 39, 1641-1653.

Montgomery, D.C. (1997). Design and Analysis of Experiments. (4Ed), John Wiley & Sons.

Nelson, P.R.C., P.A. Taylor and J.F. MacGregor (1996). Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations. Chemometrics and Intelligent Laboratory Systems, 35, 45-65.

Prasad, V.V., M. Schley, L. P. Russo and B.W. Bequette (2002). Product property and production rate control of styrene polymerization. J. Proc. Cont., 12, 353–37.

White, D.C. (2003). Creating the smart plant. Hydrocarbon Processing, October, 41-50.

Yoo, C.K. and I.B. Lee (2004). Soft Sensor and Adaptive Model-Based Dissolved Oxygen Control for Biological Wastewater Treatment Processes. Environmental Engineering Science, 21, 331-340.

Yoo, C.K., J.M. Lee, I.B. Lee and P.A. Vanrolleghem (2004). Dynamic monitoring system for full-scale wastewater treatment plants. Water Science & Technology, 50, 163-171.

Zhang, H., Z. Zouaoui and B. Lennox (2005). A Comparative Study of Soft-sensing Methods for Fed-batch Fermentation Processes. IFAC World Congress.