Korean J. Chem. Eng., **30**(1), 11-19 (2013)
DOI: 10.1007/s11814-012-0107-z

Statistical data modeling based on partial least squares: Application to melt index predictions in high density polyethylene processes to achieve energy-saving operation

Faisal Ahmed*, Lae-Hyun Kim, and Yeong-Koo Yeo*,†**

*Department of Chemical Engineering, Hanyang University, Seoul 133-791, Korea
**Department of Chemical Engineering, Seoul National University of Science and Technology, Seoul 135-743, Korea

(*Received 4 October 2011 • accepted 5 July 2012*)

Abstract−We propose two parameter update schemes which employ recursive update of partial Least Squares (PLS) model parameters as well as a model bias update to the process data. These update schemes have been applied to the successful prediction of Melt Index (MI) in grade-change operations of High Density Polyethylene (HDPE) plants. The lack of sophisticated software support hinders the recurrent use of these techniques. This paper also presents user-friendly, easy to use, graphical user interface to raise the usability and accessibility of the approach of online update of the PLS models.

Key words: PLS Model Parameters Update, Model Bias Update, HDPE, Melt Index (MI), GUI

INTRODUCTION

Accurate prediction of quality variables is essential for efficient and professional monitoring and control of chemical processes. However, due to complexity of method and time delay in measuring the quality variable analytically, inferential models have been widely applied. First principle modeling (i.e., mathematical modeling) has been successfully applied to various chemical processes including polymerization processes [1,2]. In fact, the development of the first principle model for some highly complex nonlinear processes is practically infeasible. In chemical industries, there is a huge amount of production data that possesses statistical correlations between process and quality variables that can be exploited for the prediction of the quality variables. The models developed for these relations are based on black box modeling techniques including neural networks [3], statistical data modeling (SDM) such as partial least squares regressions [4], support vector machines [5] and hybrid modeling [3,6,7]. In this era, data based modeling is now a well-developed area of research that has provided industries with various soft sensors for the prediction of quality variables. Chemical industries use these soft sensors as essential industrial equipment for efficient online monitoring of quality variables to reduce the amount of off-specification products and operation cost. Under the category of SDM, the partial least squares (PLS) technique has been a proficient and powerful multivariate regression technique for addressing noisy and highly correlated process variables [8]. It has an advantage over other methods because unlike other methods it models the input space as well as the output space, reducing the dimensionality. It simply copes with the ill-conditioning problem of ordinary least squares (OLS) by projecting the data information onto a subspace of orthogonal latent variables and regressing the input and output data by univariate regression correspondingly [4,9].

Many chemical processes, particularly polymerization processes, involve high dimensionality, collinearity, nonlinearity and grade-changes. High dimensionality and collinearity problems can be overcome by developing the linear relationship between latent vectors of the process and quality variable employing linear PLS framework. To capture the nonlinearity, many nonlinear techniques have been proposed both within and without the PLS framework. Techniques within the PLS framework include Quadratic PLS [9,10], Spline-PLS [11], neural network PLS [6,12,13] and Fuzzy-PLS [7]. Techniques without the PLS framework include the support vector machines [5], neural network model [13] and the neuro-fuzzy model [3,14]. In addition to capturing nonlinearity, the time varying nature of a process has been addressed [15-17].

Market conditions and increasing demand of products encourage frequent grade changes in the highly nonlinear, time-variant processes in the same reactor. The requirement to reduce off-spec product during grade changing events while satisfying other process constraints recommends online adaptation of the model. Time-variant processes may be modeled including process dynamics in the PLS model. However, the dynamic PLS is mostly used for the short sampling frequency processes. For processes having large offline sampling interval, dynamic PLS may not adequately capture a time-variant process [18]. Recursively adaptive data models can be employed to properly cope with the grade-changing characteristics of the HDPE process. Many researchers have paid considerable attention to online adaptation of the model based on PLS [17,18]. A recursive PLS-based soft sensor for prediction of the melt index (MI) has successfully applied to estimate MI in HDPE plant [19]. Dynamic estimation of MI based on correlation relations has also been proposed [20-22]. In this work, inspired by recursive update of PLS parameters by new process data points with mean and variance update [18] and model bias update [23], two update schemes are developed for the prediction of quality variable by the combination of these update methods. The proposed update schemes with inherent selection criteria keep track of grade-changing characteristics by model bias update. These schemes take the benefits of the recursively updated PLS by adding new process data point(s) and removing

†To whom correspondence should be addressed.
E-mail: ykyeo@hanyang.ac.kr

the oldest one(s) to update the model recursively with the update of mean and variance each time the PLS model is selected to update the model. These schemes along with comparative strategies (to be described in subsequent sections) have been employed in the graphical user interface (GUI). The GUI includes the options for the k-cross validation (CV) for the selection of number of latent factors (LF) to be used in the PLS model and optimization of parameters of different strategies and the proposed schemes. The article includes description of model update methods, the proposed schemes and their parameter's optimization and the application to the HDPE process.

BACKGROUND

1. Partial Least Squares (PLS) Method

PLS is a widely used multivariate statistical tool in chemometrics practices to handle high dimensionality and collinearity by projecting the input variable matrix **X** and output variable(s) **Y** onto the low dimensional latent subspace. Several algorithms have been proposed for the projection of input variables to the low dimensional subspace through different iterative manners [4,16]. Among them, non-iterative partial least squares (NIPALS) is the most intuitive method that calculates the outer model parameters by decomposing matrices **X** and **Y** into bilinear terms as follows [4]:

$$\mathbf{X}=\mathbf{t}_1\mathbf{p}_1^T+\mathbf{E}_1 \tag{1}$$

$$\mathbf{Y}=\mathbf{u}_1\mathbf{q}_1^T+\mathbf{F}_1 \tag{2}$$

where $\mathbf{t}_1$ and $\mathbf{u}_1$ are the latent score vectors and $\mathbf{p}_1$ and $\mathbf{q}_1$ are the loading vectors of the first latent factor of **X** and **Y**, respectively. $\mathbf{E}_1$ and $\mathbf{F}_1$ are the residuals which are to be minimized so as to extract possible maximum information from the first factor. Score vectors $\mathbf{t}_1$ and $\mathbf{u}_1$ are related to each other by a linear relation forming an inner model defined as:

$$\mathbf{u}_1=\mathrm{b}_1\mathbf{t}_1+\mathbf{r}_1 \tag{3}$$

where $\mathbf{r}_1$ is the residual which is minimized to determine the regression coefficient b_1 for the first latent factor. Further latent factors are extracted from the deflated **X** and **Y** matrices until all the required information is pulled out by the specified number of latent factors. **X** and **Y** matrices are deflated as follows:

$$\mathbf{E}_1=\mathbf{X}-\mathbf{t}_1\mathbf{p}_1^T \tag{4}$$

$$\mathbf{F}_1=\mathbf{Y}-\mathrm{b}_1\mathbf{t}_1\mathbf{q}_1^T \tag{5}$$

After extracting all the specified latent factors, PLS regression coefficient matrix is calculated as

$$\mathbf{C}_{pls}=\mathbf{W}^*\mathbf{B}\mathbf{Q}^T \tag{6}$$

where

$$\mathbf{B}=\mathrm{diag}(\mathrm{b}_1, \mathrm{b}_2, \ldots \mathrm{b}_a),\ \mathbf{W}^*=[\mathbf{w}_1^*, \mathbf{w}_2^*, \ldots, \mathbf{w}_a^*]$$

$$(\mathbf{w}_1^*=\mathbf{w}_1),\ \mathbf{w}_a^*=\prod_{k=1}^{a-1}(\mathbf{I}_K-\mathbf{w}_k\mathbf{p}_k^T)\mathbf{w}_a$$

k=(2, …a), and $\mathbf{I}_K$ is the identity matrix of dimension K. Details on NIPALS algorithm can be found elsewhere [4,17]. Usually the number of latent factors is calculated by cross validation. The root mean squared error of cross validation (RMSECV) is computed for different number of latent factors using leave-one-out (LOOCV), repeated random sub-sampling validation or V-fold cross validation (also called k-fold cross validation), while the number of optimal factors giving the least RMSECV is chosen for information extraction from input and output data set [4,24]. Leave-one-out cross validation may not be suitable for estimating generalization error of grade-changing processes because it may affect the RMSECV adversely based on the grade-changing instance. The problem is the lack of continuity of the same grade and a small change in the data can cause a large change in the model selected [25]. The same problem is with the repeated random sub-sampling which selects the test set randomly and may spread the data samples of the different grade. This sparseness of the data samples of different grade may behave as outliers, consequently giving biased result. V-fold cross validation covers this problem up to some extent by distributing the data set into V partitions and computing RMSECV by taking all the partitions as the test set one by one in rotation, remaining partitions being the training set. It was found that 10-fold and 5-fold cross validation give better results than those of LOOCV [26].

2. PLS Model Parameters Update

To address the time-varying effects of processes, the employment of an adaptive model is indispensable. A recursive PLS model was proposed by updating the training data set recursively and removing the oldest data sample(s) simultaneously [15]. The proposed recursive PLS was further extended [17]. In this way, the size of the matrices can be kept constant, the model can be adapted with new events, and the process history can be retained partially. Each new measurement added to the data set removes the oldest measurement and the PLS parameters are updated to be compatible for predicting the new process environment.

3. Mean and Variance Update

Before a model is developed, process data are often scaled to mean-center and unit variance. Mean of a given variable is subtracted from each data point of that variable followed by the division by standard deviation of same variable.

$$\mathrm{x}_{i,ms}=\frac{\mathrm{x}_i-\mathrm{m}}{\mathrm{s}} \tag{7}$$

where $\mathrm{x}_{i,ms}$ is the transformed value of x_i, i=1, 2, …N, and m and s represent the mean and the standard deviation of the corresponding variable, respectively. Whenever the model is updated by PLS on the availability of new process data, the mean and the variance are updated as well to adapt with the scaling of new process data. The updating method is given as [18]:

$$\mathrm{m}_{h+1}=\frac{\mathrm{N}-1}{\mathrm{N}}\mathrm{m}_h+\frac{1}{\mathrm{N}}\mathrm{x}_{h+1} \tag{8}$$

$$\mathrm{s}_{h+1}^2=\frac{\mathrm{N}-2}{\mathrm{N}-1}\mathrm{s}_h^2+\frac{1}{\mathrm{N}-1}(\mathrm{x}_{h+1}-\mathrm{m}_{h+1})^2 \tag{9}$$

where m_h and s_h are the mean and the variance of training data at the h^{th} addition of the new measurement, respectively, and m_{h+1} and s_{h+1} represent the corresponding values at $(\mathrm{h}+1)^{th}$ addition.

4. Model Bias Update

Drifts in process environment with time may affect the relation between process input variables and output quality variable. Moreover, grade-changing events in the operations of chemical processes may result in PLS overfitting the data and giving deviated and undesirable predictions. To circumvent this situation, model bias update is incorporated in the soft sensor to make it reliable and robustly adaptive [23]. The difference of predicted value and corresponding measurement, calculated by Eq. (11), is termed as model bias which is added to the predicted value at the next time step to modify the prediction. At t=0,

$$\mathbf{Y}_{pred}=\mathbf{X}(t)\times\mathbf{C}_{pls} \tag{10}$$

where t is the index of model update run: t=0 represents the instance before the update. At t=tth run,

$$\text{bias}(t)=\mathbf{Y}_{lab}(t-1)-\mathbf{Y}_{pred}(t-1) \tag{11}$$

where bias(t) is the model bias at the tth run to be used in the model output modification by Eq. (12) with bias(0)=0.

$$\mathbf{Y}_{mod}(t)=\mathbf{Y}_{pred}(t)+\text{bias}(t) \tag{12}$$

where $\mathbf{Y}_{lab}$ and $\mathbf{Y}_{pred}$ represent the measurement and the predicted value respectively. $\mathbf{Y}_{mod}$ is the modified value of $\mathbf{Y}_{pred}$ by the model bias.

NEW UPDATE SCHEMES

The proposed update schemes consist of the model adaptation with the recursive update of the parameters of the PLS model and the model bias update as well. The main idea of the schemes is to combine the two update methods and devise a selection criterion to choose one method for a specific instance of update in a fashion so as to minimize the prediction error. The model is updated with the update methods one at a time, i.e., the PLS model parameters update or the model bias update. For the performance evaluation, relative RMSE given by Eq. (13) is used:

$$\text{RMSE}=\sqrt{\frac{1}{\text{N}_t}\sum_{i=1}^{N_t}\left(\frac{\mathbf{Y}_{i,actual}-\mathbf{Y}_{i,mold}}{\mathbf{Y}_{i,actual}}\right)^2} \tag{13}$$

where N_t is the number of observations in the test data set whereas $\text{Y}_{i,\,actual}$ and $\text{Y}_{i,\,mod}$ are the actual and modified values of the MI. The term RMSE used in the article represents the relative RMSE.

1. Proposed Scheme-I

In most industrial processes, it takes quite a long time (often several hours) to measure the quality variable. During the first interval after the training data is collected and the initial model is built, quality variable is predicted through the model initially built by the training data samples. Later on the PLS model parameters or model bias is updated on the arrival of each new measurement and predictions are carried out for the next interval. The selection criterion of the update methods for this scheme are based on the threshold constant d below which there is no effect on the relative root mean square error (RMSE) of predictions. This threshold constant decides whether to use PLS model parameters update or model bias update depending on the changing behavior of quality variable. Threshold constant is an arbitrary constant which is optimized for a certain set of calibration data (see section 2.3). The absolute value of the difference between the values of current and previous measurements is termed as adiff. When a new measurement is available, adiff is calculated and compared with the threshold constant d. If the adiff is less than or equal to d, PLS model parameters update is selected, otherwise the model is updated by model bias update. At first we intended to use the PLS model parameters update during small variations in quality variable behavior, and model bias update to capture a large impulse (grade changing event). However, smaller values of d affect the model predictions positively at the occurrence of undersized variations along with the grade changing events. The update of PLS model parameters is performed followed by mean and variance update until adiff reached the threshold constant or the process changes its grade. With a value of adiff higher than d, the model bias term is updated to new value. Eventually, all the predictions are modified by the model bias using Eq. (12). The procedure is illustrated in Fig. 1.

When the process shifts from one grade to another, the difference of the first value of the shifted grade and immediately previous value belonging to the prior grade is termed as grade-change-defining-value (GCDV). When the operation consists of several grades, GCDV may be different for various grades. In this case, the minimum GCDV is selected for the optimization, which is 1.94 for this application. The threshold constant is obtained by the optimization; one can start with the GCDV proceeding towards lower

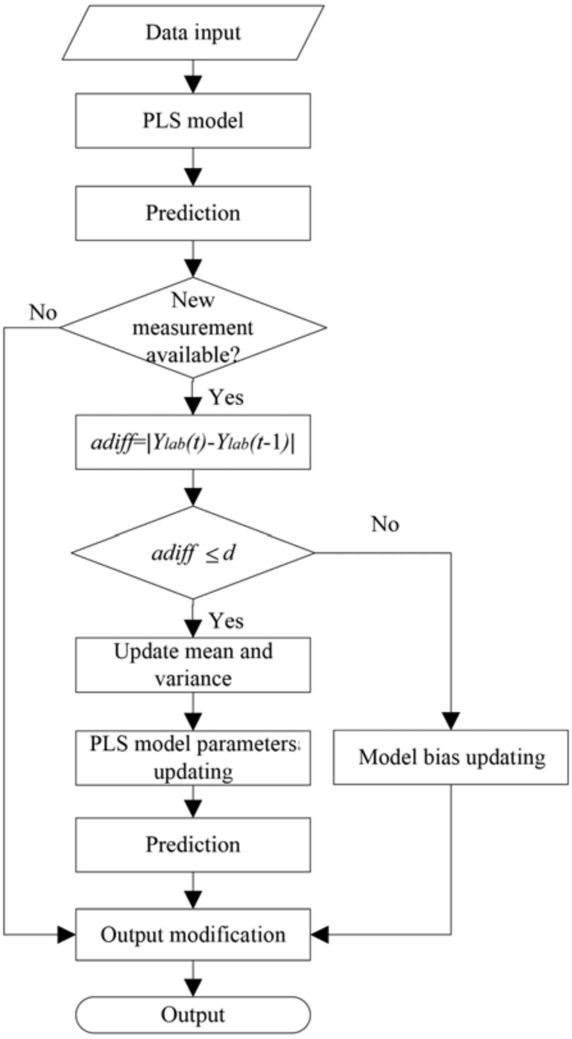


Fig. 1. Flowchart of the update scheme-I.

values. The threshold value giving the lowest RMSE is used as the update method selector d (see section 2.3).

2. Proposed Scheme-II

Despite the fact that the model adaptation is vital for the soft sensor to predict deviated response of the process, undue updating may encompass potential to cause relative overfitting. Proposed scheme-II is the extension of scheme-I. It detects the process response behavior more critically and is developed with the aim of making the soft sensor more sensitive for the selection of update method to prevent undue model updating. This prevention from unnecessary updating circumvents from the potential risk of overfitting and over modification, on the other hand, automatically reduces the computation time as well up to some extent.

During update selection procedures preference is given to no update at all (when there is no need to update) followed by model bias update and PLS model parameters update. In other words, the parameters are to be optimized so as to minimize the number of PLS model parameters update runs (NPR). A lower bound value d_1 is identified below which the system, if left without updating, does not exhibit any remarkable change and has no or negligible increasing effect on the relative RMSE. Therefore, the model does not require updating in these cases. On the other hand, to capture the sharp and rapid changes an upper bound d_2 is selected so that the model bias update updates the bias above d_2. The value of d_2 is optimized to make use of the model bias update during the significant changes within the same grade and drastic changes of grade-changing events as well. The range (the difference between lower and upper threshold bounds) represents the activation of the RPLS update.

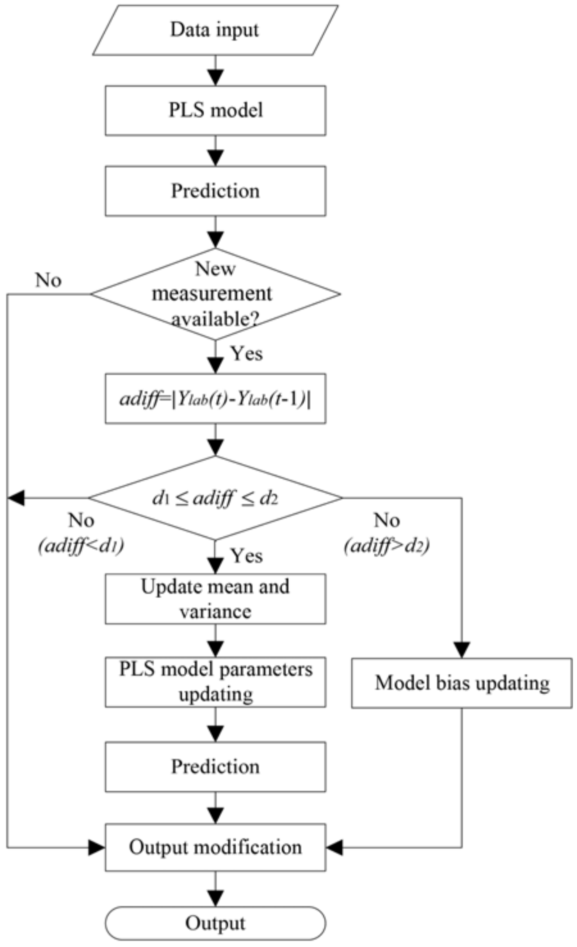


Fig. 2. Flowchart of the update scheme-II.

Scheme-II follows the same procedure as that of scheme-I except the selection criteria. If the value of adiff lies out of the ranges of d_1 or d_2, no update or the model bias update is performed. Otherwise, the PLS model parameters update is activated. Eventually, as in scheme-I, predictions from both update methods are modified by the model bias using Eq. (12). The procedure is illustrated in Fig. 2.

Scheme-II is set up with the minimized NPR, which may cause a negligible increase in RMSE. The value up to which the increase in RMSE is acceptable is termed as the compensation factor (cf). The placement and the range between lower and upper bounds are to be optimized to minimize the NPR. One can start with d (parameter of scheme-I, threshold constant) and far less value than GCDV as the lower and upper bounds, respectively, proceeding towards each other. GCDV is the maximum limit for the upper bound but usually the optimization results in upper bound value near to lower bound value and far less than GCDV. The placement of the bounds and the range giving the least RMSE are used as the update method selection criteria as described in the subsequent section.

3. Determination of the Threshold Constant and Bounds

For the determination of the parameters for scheme-I, the problem can be set up as the unconstrained optimization problem with the RMSE as the objective function to be minimized subject to d.

Minimize: RMSE(d)=scheme-I (14.1)

Subject to: $0 \leq d \leq \mathrm{GCDV}$ (14.2)

The threshold constant d turns out to be a variable to be optimized and is related to the RMSE through the update scheme-I. A value of d closer to GCDV results in frequently activated PLS model parameters update, which leads towards the potential of overfitting. On the other hand, the model bias update is activated too frequently for the value of d nearly zero, resulting in the possibility of over modification.

For scheme-II, the determination of the parameters is a constrained optimization problem with RMSE, NPR as objective function to be minimized, and bounds placement and range as the variables to be optimized. The optimization problem takes the form as:

Minimize: NPR=scheme-II (15.1)

Subject to: RMsE II≤RMSE I+cf (15.2)

$0 \leq d_1 \leq \mathrm{GCDV};\ d < d_2 \leq \mathrm{GCDV}$ (15.3)

where RMSE I and RMSE II represent the RMSE obtained by the scheme-I and scheme-II, respectively and cf is the compensation factor. If the values of d_1 and d_2 move simultaneously in the same direction on a scale from zero to GCDV, they can be considered as a single parameter which affects the selection of update method by moving towards left minimizing the probability of selecting no update and towards right maximizing the probability of selecting no update. The value of d_1 and d_2 approaching or going farther from each other decides the range which corresponds to the selection of PLS model parameters update.

A graphical user interface (GUI) can be easily constructed for convenient use of the proposed model adaptation schemes. The GUI shown in Fig. 3 is developed with the compatibility with the Microsoft

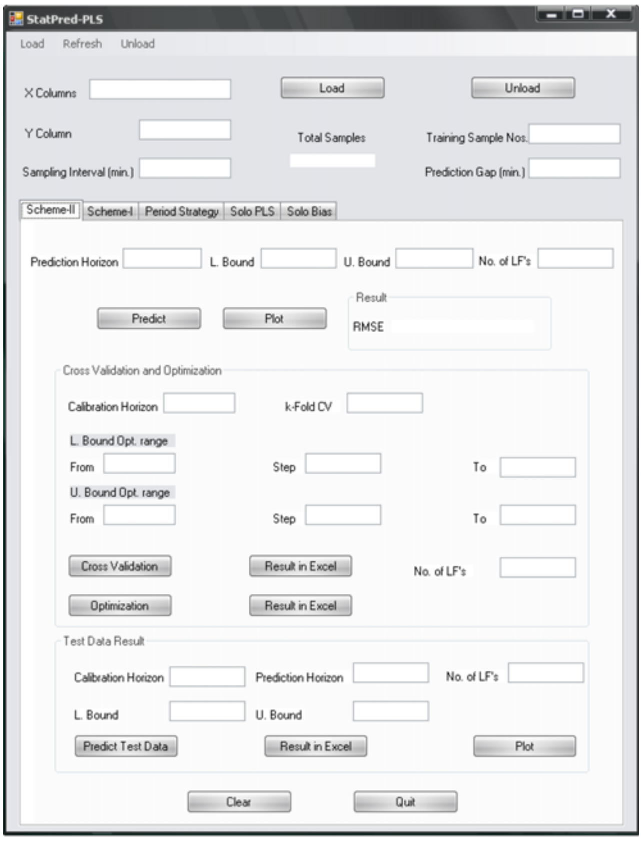


Fig. 3. A snapshot of StatPred-PLS.

Excel and Matlab and mainly consists of two sections. The first section is a common input section in which some input fields appear that are common to all strategies and schemes. The other section consists of five different tabs; each tab has input fields for distinct specific input parameters for prediction, cross validation and optimization, and buttons for displaying output in MS Excel sheet. Both the solo strategies are the exceptions; these two strategies do not have any parameters that require optimization. The *Load* button opens a dialogue box asking for an excel file. Once the file is selected it detects the number of sheets existing in the excel file and extracts from all sheets the columns as inserted in the *X Columns* and *Y Column* input fields. Based on the sampling interval inserted in the respective field, GUI displays the total number of samples available for model fitting, calibration and/or testing. Pressing the *Load* button and selecting a file (same or other), while the previous data is already loaded, concatenates newly loaded data to the bottom of previous data. This concatenation can be done for more than two files. The number of columns extracted from all files must be same for the concatenation. After the other file is loaded, the *Total Samples* field shows increased number of samples.

Under each tab, the strategies and schemes have distinct parameters but the method of using the GUI is the same. We will discuss the *Scheme-II* tab. Under the *Scheme-II* tab, *Predict* button runs the scheme-II algorithm in Matlab and displays the relative RMSE on the right. These buttons are used when one knows the number of latent factors and the optimized parameters. Matlab's graph appears showing the comparison between original and predicted values of quality variable on pressing the *Plot* button. Under the *Cross Validation* and *Optimization* subsection, *Lower Bound Opt. range* and *Upper Bound Opt. range* are the input ranges for the optimization of the parameters. *Cross Validation* and *Optimization* buttons run the optimization algorithm based on scheme-II and k-fold cross validation algorithm, respectively. *Result in Excel* buttons displays the corresponding result in Excel sheet. The subsection *Test Data Result* is used after the number of latent factors and parameters are optimized by the *Cross Validation and Optimization* subsection. The *Predict Test Data* button initializes the predictions from the calibration data and continues to predict the test data, but the result and plot are for test data only. The *Predict* button does not include the calibration data in the algorithm to initialize predictions. The *Unload* button simply unloads the loaded data.

APPLICATION TO THE HDPE PROCESS

1. Process Description

HDPE is produced under the edge-cutting low-pressure polymerization manufacturing process in LG Petrochemicals plant located in the southwestern region of Korea. There are two polymerization processes named K_1 and K_2 in the plant. Two parallel reactors are employed in the K_1 process, whereas the K_2 process uses a cascade arrangement. Due to the exothermic nature, these reactions generate 1,000 kcal/kg ethylene. To remove polymerization heat from the reactor, an efficient cooling system is used. The reactant feed of the reactor includes ethylene co-monomer, hydrogen, activator, catalyst, co-catalyst, and hexane and continuously recycled mother liquor. The reactor volume is filled up to 90-95% with reaction slurry, which is transferred to the subsequent equipment with the rise in pressure to maintain the slurry level within the reactor. The pressure ranges for K_1 and K_2 processes operate under 8-10 kg/cm^2 and 2-4 kg/cm^2 respectively with a temperature range of 74-85 ^{o}C in both processes.

2. Numerical Simulations

For the simulation purpose, 1156 measurements were collected with the interval of 2 hours in 97 days. Out of 1156, first three hundred measurements were selected for training of the initial model. For calibration of model parameters through optimization, 206 measurements were used from which one measurement was used for the predictions before starting the online update, and 650 measurements were used as test data to examine the model for prediction accuracy. Variable selection is a crucial task in developing a statistical model. In the LG Petrochemicals plant, 43 input variables are recorded on a daily basis. Activator feed rate was observed constant throughout the process, so it was deleted from the data. From the remaining 42 process input variables, 14 variables were selected using interval PLS for the input variable selection with the help of PLS_Toolbox 4.2, Eigenvector Research Inc. An initial PLS model was set up with 300 data points for the update strategies, and proposed schemes using one latent factor except for solo PLS strategy for which five latent factors were used. The initially built model was used for the predictions with a gap of 5 minutes before the model started online adaptation until the entrance of new measurement in the input data. The time required by the offline measurements of MI in LG petrochemicals is two hours. After each two hours at the arrival of a new measurement of MI, it is added to the existing process data and a certain update method is activated to update the PLS model parameters or model bias. The recently updated model then

Table 1. CV Results: RMSECV for strategies and proposed schemes against LF

LF	Scheme-II	Scheme-I	Period strategy	Solo PLS strategy	Solo bias strategy
1	1.2259	1.2478	1.2517	4.5624	1.1105
2	1.6291	1.2271	1.2419	4.5297	1.5986
3	1.8127	1.7379	1.3801	2.9777	1.7369
4	1.9448	1.7420	1.4691	1.9082	1.6580
5	1.6850	1.7905	1.4499	1.7000	1.4337
6	1.8739	1.7952	1.4436	1.5942	1.6874
7	1.7654	1.7491	1.4041	1.5585	1.5169
8	1.4310	1.7618	1.4058	1.5370	1.3387
9	1.3913	1.7313	1.4107	1.4983	1.2794
10	1.3463	1.6925	1.3979	1.4601	1.2118
11	1.4927	1.7334	1.4110	1.4493	1.3300
12	1.6493	1.7424	1.4156	1.4156	1.4824
13	1.2172	1.7790	1.4081	1.4146	1.1080
14	1.0978	1.7560	1.4124	1.3846	1.0055

predicts MI for next two hours unless a new measurement is carried out analytically in lab. In this fashion, model update and predictions are executed sequentially with the interval of two hours. The number of latent factors to be used in PLS model for update strategies and proposed schemes was selected using 5-fold cross validation. CV results for strategies and the proposed schemes are tabulated in Table 1. Although the 14 latent factors give the least RMSECV but the difference in RMSECVs against 1 latent factor and 14 latent factors is not considerable. It clearly suggests one latent factor for scheme-II. Similarly, one latent factor was selected for scheme-I, period strategy and solo bias strategy. For the solo PLS strategy an optimal number of latent factors equal to five was selected. The different strategies and proposed schemes are set up as follows:

2-1. Solo PLS Strategy

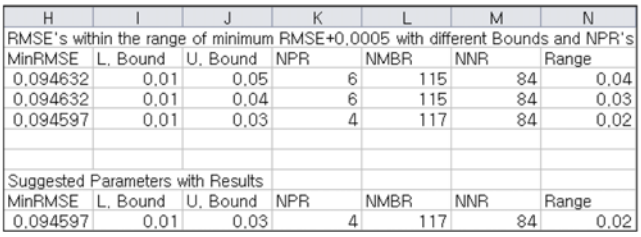
At each arrival of a new MI measurement in the training data set, the parameters of the PLS model are updated recursively along with means and variances which are updated by Eqs. (8) and (9). There is no selection of model bias update or no update in this strategy. The PLS model then uses the updated parameters to predict the MI values for the interval. For the PLS model five latent factors were selected based on the cross validation result from GUI as shown in Table 1.

2-2. Solo Bias Strategy

After the initial predictions for the first interval through the initial PLS model, solo bias strategy updates the model bias at the arrival of each new measurement. For the subsequent intervals the model bias is added to the model predictions to incorporate the bias effect into the predictions. One latent factor was used for the initial PLS model.

2-3. Period Strategy

It utilizes the arbitrarily constant parameter *period* throughout the model predictions. *Period* equal to two acts as a switch between PLS model parameters update and model bias update, and activates them alternately. Similarly, *period* equal to three updates the bias every two consecutive intervals and PLS model parameters every third interval. One latent factor was selected for the initial and subsequent PLS update models. Optimization results generated by the GUI suggested *period*=7 giving the least RMSE for the calibration data.



H	I	J	K	L	M	N
RMSE's within the range of minimum RMSE+0,0005 with different Bounds and NPR's						
MinRMSE	L, Bound	U, Bound	NPR	NMBR	NNR	Range
0,094632	0,01	0,05	6	115	84	0,04
0,094632	0,01	0,04	6	115	84	0,03
0,094597	0,01	0,03	4	117	84	0,02
Suggested Parameters with Results						
MinRMSE	L, Bound	U, Bound	NPR	NMBR	NNR	Range
0,094597	0,01	0,03	4	117	84	0,02

Fig. 4. Scheme-II: Cut off results of optimization through GUI shown in microsoft excel.

2-4. Proposed Scheme-I

For the proposed scheme-I, optimization of threshold constant was carried out by the GUI using the calibration data resulting in the value of d equal to 0.01 in this application, and latent factor equal to one based on the result of CV.

2-5. Proposed Scheme-II

The proposed scheme-II is characterized by the lower and upper bounds d_1 and d_2, which were optimized using calibration data giving d_1 and d_2 equal to 0.01 and 0.03 respectively as shown in Fig. 4. One latent factor was used for the PLS model.

RESULTS AND DISCUSSION

1. Results

Results of predictive ability of the strategies and proposed schemes are compared in Fig. 5. The solo PLS strategy and solo bias strategy

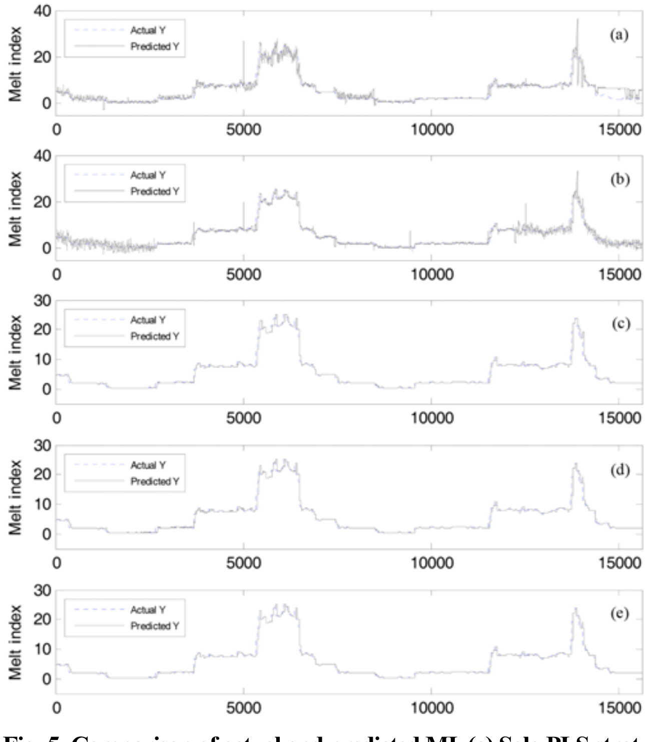


Fig. 5. Comparison of actual and predicted MI. (a) Solo PLS strategy; (b) Solo bias strategy; (c) Period strategy with *period* =7; (d) Proposed scheme-I; (e) Proposed scheme-II.

capture the grade-changing events but with over modification at some instances giving noise and shoot-ups in predictions. The solo PLS strategy gives the predictions up to the acceptable range, except for some instances where it produces spikes and near the end where it loses the trend (Fig. 5(a)). Solo bias strategy tends to over-modify the model predictions, giving some spikes as shown in Fig. 5(b). The reason for the spikes may be the noise and/or a change in the process input variables at the instances near or during the grade-changing events. The noise and a change in process input variables have the tendency for the update method to over-modify the predictions, and solo strategies are vulnerable to this tendency. In contrast to the solo strategies, period strategy (Fig. 5(c)) circumvents the situation of over modification imposed by the noise or a change in input variables. It predicts the MI smoothly and overcomes the deviations in predictions by activating the PLS at a *period* equal to seven. Fig. 5(d) represents scheme-I, which is characterized by its sophisticated ability to choose an appropriate update method at a certain instance outperforms and predicts the MI more closer to the actual values of MI giving lower RMSE than that given by the period strategy. Proposed scheme-II (Fig. 5(e)) minimizes the NPR by optimizing the bounds placement and limiting the range between d_1 and d_2 and minimizes the RMSE as compared to the RMSE from scheme-I. Values of RMSE of calibration and test data for the strategies and schemes are shown in Table 2.

Table 2. Comparison of update strategies and schemes

No.	Procedure	RMSE (calibration data)	RMSE (test data)
1	Solo PLS strategy	0.3611	0.6418
2	Solo bias strategy	0.5209	0.7569
3	Period strategy	0.0961	0.1533
4	Scheme-I	0.0943	0.1504
5	Scheme-II	0.0946	0.1466

The optimization was performed for both schemes with calibration data using Eqs. (14) and (15). The results were truncated to focus on the small variations in the minimum RMSE value as shown in Figs. 6 and 7. Therefore, the iteration numbers shown in Figs. 6 and 7 and given in Table 3 are not the same as the actual iteration numbers of the performed optimization. The optimization for the scheme-I converges at the 40*th* iteration giving RMSE=0.0943 at d=0.01 with the NPR=84 as shown in Fig. 6. Utilizing an analogous approach as above, the optimization was performed for scheme-II, giving variation in NPR and RMSE with different bounds placements and ranges shown in Fig. 7. The minimized NPR=4 with acceptable compensation (cf=0.0003 in this case) in RMSE was found at d_1=0.01 and d_2=0.03 (8th iteration). The results of optimization for the period strategy and the schemes are summarized in Table 3. It also tabulates the results of comparison between the NPR of the period strategy and that of proposed schemes using calibration data set and test data set as well. In Table 3, the sum of NPR, NMBR and NNP for calibration data is 205 because the first measurement is used for the prediction by initial model without updating.

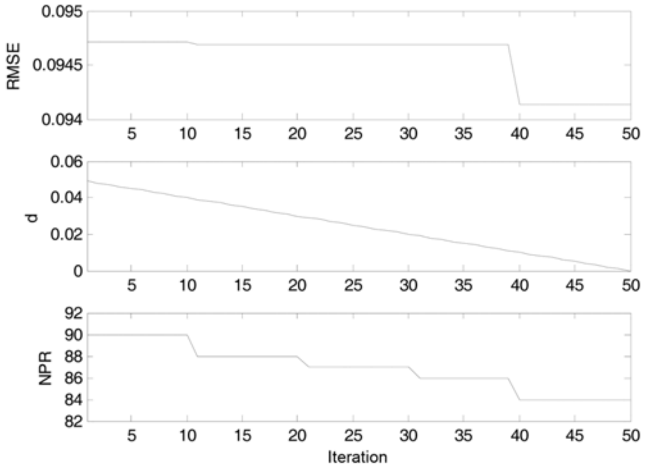


Fig. 6. Threshold constants corresponding to values of RMSE and NPR.

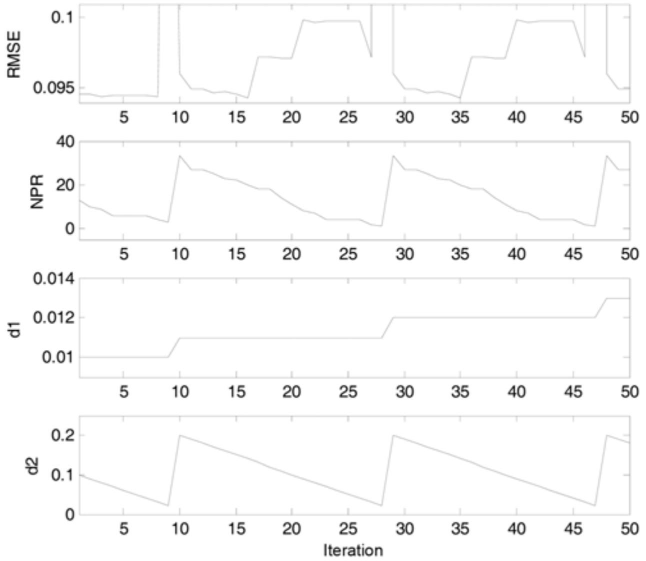


Fig. 7. Bounds placements and ranges corresponding to the values of NPR and RMSE.

Table 3. Parameter comparison among update schemes and period strategy

Update scheme	NPR	NMBR	NNP	Optimized parameters	Iteration number	NPR	NMBR	NNP
Period strategy	29	176	-	*period*=7	-	93	557	-
Scheme-I	84	121	-	d=0.01	40th (Fig. 6)	289	361	-
Scheme-II	4	117	84	d_1=0.01, d_2=0.03	8th (Fig. 7)	39	322	289

NMBR=number of model bias update runs
NNP=number of no update runs
Last three columns refer to test data set

2. Discussion

Continuous polymerization processes are characterized by frequent grade changes, and for that purpose the process input variables are to be changed to shift to the other grade. The solo PLS strategy renovates the PLS parameters recursively, overcoming the deviation caused by the abrupt changes in process data and capturing the trend appropriately except for giving some spikes at instances near the grade change where it over-fits the newly available data, but soon after captures the trend again. Solo bias strategy is also susceptible to these changes and over-modifies the predictions at these instances by updating the model bias unduly. It depicts the input data noise in the predictions up to a relatively large extent caused by the absence of PLS model parameters update. In fact, the PLS method models the output space as well as the input space, creating better inner relation between input and output score vectors. Subsequently, the solo PLS strategy, even at the instances of noisy input data, gives less noisy predictions relative to that of model bias strategy. Hence, the combination of both the methods is exploited in the period strategy and proposed schemes. The period strategy circumvents the problem of over-modification and contends with the rapid changes by using the PLS model parameters update and model bias update activated at the optimized *period*. The PLS model parameters update and the model bias update complement each other, giving lower RMSE of predictions by reducing the risk of over-fitting and over-modification. The proposed schemes exploit this combination with an inherent and decisive selection criterion that makes the schemes more sophisticated.

The proposed schemes are not only found robust in predictions without over-fitting and over-modification but also outperform the other strategies. The reason for superiority may be described as: during grade-changing industrial operations, the instances or the time difference between two consecutive grade-changing events may differ. Furthermore, the duration for which the product of one grade is manufactured may also be changed. An arbitrarily fixed parameter "*period*" is used in period strategy throughout the model and the calibration of specific period of grade changing instances is used for the optimization of *period* for the lowest RMSE. On the other hand, the proposed scheme-I shifts the decision power of selection from being based on calibration data to online quality variable behavior and selects the update method based on the current requirement of the process. In addition, scheme-II further adds the possibility of no update at certain instances. These capabilities render the selection criterion of scheme-II more sensitive for the detection of process response behavior that helps in selecting the appropriate update method circumspectly, where no, undersized or relatively large variations are identified during the operation and are treated with no update, the PLS model parameters update or the model bias update respectively.

CONCLUSIONS

New update method selection criteria for two different update methods are introduced. The proposed update method selection criteria are used in two different schemes for the prediction of MI values of HDPE with grade changing behavior. Process modeling through scheme-I focuses on the minimization of RMSE by taking advantage of activating two update methods, PLS model parameters update and model bias update, with the online selection criterion called threshold constant. Scheme-II centers on the idea of minimizing the NPR while maintaining the RMSE with some compensation factor and in some cases even minimizing the RMSE with respect to that of resulted from scheme-I. The update method selection criteria employed in proposed schemes render them able to cope with the irregular grade-changing events along with regular operation. We are looking forward to having these schemes being employed in the soft sensors categorized by online model adaptation.

ACKNOWLEDGEMENT

This work was supported by a Korea Research Foundation Grant funded by the Korean Government (2010-0007152) and in part by the Ministry of Knowledge Economy, Republic of Korea, as a part of the research project titled "Constitution of energy network using District heating energy" (Project No: 2007-E-ID25-P-02-0-000). The authors wish to thank them for their support.

NOMENCLATURE

a	: index of factors (a=1, 2…, A) [-]
A	: number of factors in PLS model [-]
b	: inner model coefficient [-]
B	: matrix of regression coefficients, size (K*M) [-]
$\mathbf{C}_{pls}$	: PLS regression coefficient matrix [-]
d	: threshold constant [-]
d_1	: lower bound [-]
d_2	: upper bound [-]
E	: residual matrix for **X** [-]
F	: residual matrix for **Y** [-]
k	: index of w (k=1, 2, …, a) [-]
K	: no. of variables in **X** [-]
N	: no. of observations (samples) in training data set [-]
N_t	: no. of observations (samples) in test data set [-]
p	: loading vector for **X** [-]
q	: loading vector for **Y** [-]
Q	: weight matrix for **Y**, size (M*A) [-]
r	: residual vector for inner model of PLS [-]
range	: difference between d_2 and d_1 [-]
t	: score vector for **X** [-]
u	: score vector for **Y** [-]
w	: column vector of **W** [-]
W	: weight matrix for **X**, size (K*A) [-]
w*	: column vector of **W*** [-]
W*	: matrix of transformed PLS weights [-]
X	: matrix of process input data, size (N*K) [-]
Y	: matrix of response variable, size (N*M) [-]
$\mathbf{Y}_{mod}$	: modified value of response variable by model bias [-]
$\mathbf{Y}_{pred}$	: prediction value of response variable [-]

REFERENCES

1. K. B. McAuley and J. F. MacGregor, *AIChE J.*, **37**, 825 (1991).
2. K. B. McAuley and J. F. MacGregor, *AIChE J.*, **38**, 1564 (1992).
3. L. Jia, E. Li and J. Yu, *Eng. Appl. Artif. Intell.*, **16**, 11 (2003).
4. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, **185**, 1 (1986).

5. V. Vapnik, *The nature of statistical learning theory*, Springer Verlag, New York (1995).
6. B. Walczak and D. L. Massart, *Anal. Chim. Acta*, **331**, 177 (1996).
7. Y. H. Bang, C. K. Yoo and I.-B. Lee, *Chemom. Intell. Lab. Syst.*, **64**, 137 (2003).
8. S. Wold, M. Sjostrom and L. Eriksson, *Chemom. Intell. Lab. Syst.*, **58**, 109 (2001).
9. S. Wold, N. Kettaneh-Wold and B. Skagerberg, *Chemom. Inellt. Lab. Syst.*, **7**, 53 (1989).
10. G. Baffi, E. B. Martin and A. J. Morris, *Comput. Chem. Eng.*, **23**, 395 (1999).
11. S. Wold, *Chemom. Intell. Lab. Syst.*, **14**, 71 (1992).
12. C. Li, H. Ye, G. Wang and J. Zhang, *Chem. Eng. Technol.*, **28**(2), 141 (2005).
13. Z. Liu, X. Wang, X. Lian, Z. Wang and C. Hou, Int. Conf. on Compt. Intelligence for Modeling, Control and Automation, and on Intelligent Agents, Web Tech. and Internet Comm. (CIMCA-IAWTIC'06) (2006).
14. J. X. Luo and H. Shao, *Soft Comput.*, **10**, 54 (2006).
15. K. Helland, H. E. Berntsen, O. S. Borgen and H. Martens, *Chemom. Intell. Lab. Syst.*, **14**, 129 (1992).
16. B. S. Dayal and J. F. MacGregor, *J. Process Control*, **7**, 169 (1997).
17. S. J. Qin, *Comput. Chem. Eng.*, **22**, 503 (1998).
18. S. Mu, Y. Zeng, R. Liu, P. Wu, H. Su and J. Chu, *J. Process Control*, **16**, 557 (2006).
19. A. Faisal, N. Salman and Y. K. Yeo, *Korean J. Chem. Eng.*, **26**(1), 14 (2009).
20. A. R. Seong, E. H. Lee, K. N. Lee and Y. K. Yeo, *Korean J. Chem. Eng.*, **26**(1), 7 (2009).
21. T. Y. Kim and Y. K. Yeo, *Korean J. Chem. Eng.*, **27**(6), 1669 (2010).
22. E. H. Lee, T. Y. Kim and Y. K. Yeo, *Korean Chem. Eng. Res.*, **46**(6), 1043 (2008).
23. R. U. Sharmin, U. Sundararaj, S. Shah, L. V. Griend and Y. J. Sun, *Chem Eng. Sci.*, **61**, 6372 (2006).
24. M. Stone, *Math. Operationsforch. Statist., Ser. Statist.*, **9**(1) (1978).
25. L. Breiman, *Annals of Statistics*, **24**, 2350 (1996).
26. L. Breiman and P. Spector, *International Statistical Review*, **60**, 291 (1992).