Development of Correlation-Based Pattern Recognition Algorithm and Adaptive Soft-Sensor Design

Koichi Fujiwara*, Manabu Kano, Shinji Hasebe

Department of Chemical Engineering, Kyoto University, Nishikyo-ku, Kyoto 615-8510, Japan

Abstract

Although soft-sensors have been used for estimating product quality, they do not always function well due to not only changes in process characteristics but also the individual difference of production devices. Correlation-based Just-In-Time (CoJIT) modeling has been proposed to cope with such changes in process characteristics; however it cannot deal with the individual difference. In the present work, a new pattern recognition method, referred to as the nearest correlation (NC) method is proposed to cope with the individual difference. The proposed NC method is integrated with CoJIT modeling. The advantages of the proposed methods are demonstrated through a case study.

Key words: Soft-sensor, Process observation and parameter estimation, Just-In-Time modeling, Pattern recognition, Principal component analysis

*Corresponding author. Tel.: +81-75-383-2677; fax: +81-75-383-2657.
Email address: fujiwara@cheme.kyoto-u.ac.jp (Koichi Fujiwara)

1. Introduction

In many processes, it is rarely the case that product quality or other important variables are directly measured in real-time, since on-line accurate measurement of them is difficult. For example, most analyzers, like gas chromatographs, suffer from large measurement delays and high investment and maintenance costs. In such a case, a soft-sensor that can estimate product quality or other important variables from on-line measured variables is a key technology for realizing efficient operation. In addition, the estimates of soft-sensors can be used for control, and this control scheme is called inferential control (Kresta, Marlin, & MacGregor, 1994; Kano, Miyazaki, Hasebe, & Hashimoto, 2000).

Partial least squares (PLS) regression and artificial neural network (ANN) have been widely accepted as useful techniques for linear and nonlinear soft-sensor design (Kano, & Nakagawa, 2008; Mejdell, & Skogestad, 1991; Kamohara, et al, 2004; Radhakrishnan, & Mohamed, 2000). Recently, support vector regression (SVR) and Bayesian method have been used for soft-sensor design (Jain, Rahman, & Kulkarni, 2007; Desai, Badhe, Tambe, & Kulkarni, 2006; Khatibisepehr, & Huang, 2008). In addition, the application of subspace identification (SSID) to soft-sensor design has been reported for achieving higher estimation performance (Amirthalingam & Lee, 1999; Kano, Lee, & Hasebe, 2009).

Generally, building a high performance soft-sensor is very laborious, since input variables and samples for model construction have to be selected carefully and parameters have to be tuned appropriately. In addition, even if a good soft-sensor is developed successfully, its estimation performance dete-

riorates as process characteristics change. In chemical processes, for example, process characteristics are changed by catalyst deactivation or fouling. In semiconductor manufacturing processes, periodic cleaning of equipment changes the process characteristics dramatically. Such a situation may deteriorate product quality. Therefore, maintenance of soft-sensors is very important in practice to keep their estimation performance. Kano, & Ogawa (2009) concluded that soft-sensors should be updated as the process characteristics change, and also manual and repeated construction of them should be avoided due to its heavy workload.

To update statistical models automatically when process characteristics change, recursive methods such as recursive PLS were developed (Qin, 1998). These methods can adapt models to new operating conditions recursively. However, when a process is operated within a narrow range for a certain period of time, the model will adapt excessively and will not function in a sufficiently wide range of operating conditions. In addition, recursive methods cannot cope with abrupt changes in process characteristics, which take place in the semiconductor industry as mentioned above.

The Just-In-Time (JIT) modeling has been proposed to cope with process nonlinearity (Bontempi, Birattari, & Bersini, 1999a; Atkeson, Moore, & Schaal, 1997) and changes in process characteristics (Cheng, & Chiu, 2004). In JIT modeling, a local model is built from past data around the query only when an estimate is required. JIT modeling is useful when global modeling does not function well. However, its estimation performance is not always high because the samples used for local modeling are selected on the basis of the distance from the query and the correlation among variables is

not taken into account. How should we determine the samples used for local modeling to build a highly accurate statistical model? Distance is not the most important. A good model cannot be developed when correlation among input-output variables is weak, even if the distance between samples is very small. Conversely, a very accurate model can be developed when the correlation is strong even if the distance is large.

Recently, a new JIT modeling method based on the correlation among variables, referred to as the correlation-based JIT (CoJIT) modeling, was proposed (Fujiwara, Kano, & Hasebe, 2009). In CoJIT modeling, candidate data sets for local modeling are constructed so that they consists of successive samples included in a certain period of time by using moving time-window, and the data set is selected on the basis of correlation. CoJIT modeling can cope with abrupt changes in process characteristics and also achieve high estimation performance.

In addition, the individual difference of production devices should be taken into account. In semiconductor processes, for example, tens of parallelized production devices are used, and they have different characteristics even if their catalog specifications are the same. Therefore, a soft-sensor developed for one device is not always applicable to another device, and it is very laborious to customize soft-sensors according to their individual difference. In this case, the correlation-based method is also applicable because the individual difference of production devices is expressed as differences of the correlation among variables.

However, CoJIT modeling is applicable to only time-series data because it uses moving time-windows to generate data sets for local modeling. In

other words, CoJIT modeling cannot generate a data set consisting of such data that represent characteristics of a query and are obtained from various devices operated in parallel.

To design a soft-sensor for parallelized production devices, samples obtained from various devices have to be discriminated on the basis of the correlation among variables, since a data set that represent characteristics of a query should be constructed for local modeling. This discrimination problem is one of the unsupervised pattern recognition problems because a teacher signal is not used for sample classification.

The nearest neighbor (NN) method and the k-means method are well-known conventional unsupervised pattern recognition algorithms. The NN method can detect samples that are similar to the query, and the k-means method can cluster samples without a teacher signal. However, they are distance-based methods and do not take into account the correlation among variables. Self organizing map (SOM) is another unsupervised pattern recognition method (Kohonen, 2001). SOM is a machine learning process that imitates the brain learning process, and it can visualize high dimensional data as a map on the basis of the similarities among samples. However, SOM does not always give clear boundaries between clusters on the map. In addition, it requires high computational load, and its data preprocessing is complicated.

In the present work, to cope with the individual difference of production devices as well as changes in process characteristics, a new unsupervised pattern recognition method based on the correlation among variables, referred to as the nearest correlation (NC) method, is proposed. The proposed NC method can detect samples that have correlation similar to the query on the

basis of sample geometry. In addition, it is integrated with CoJIT modeling. The usefulness of the integration is demonstrated through a case study of a parallelized chemical reaction process.

2. Indices of Correlation

In this section, several measures for quantifying correlation among variables are briefly explained. In this manuscript, "correlation" means "relationship" among variables.

2.1. Correlation coefficient

The correlation coefficient $C_{i,j}$ can be used as an index of the similarity between two vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j \in \Re^M$.

$$C_{i,j} = \frac{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j}{||\boldsymbol{x}_i|| ||\boldsymbol{x}_j||} = \cos\theta \tag{1}$$

where θ is the angle between two vectors.

Suppose that the samples in the three-dimensional data consist of two classes K_1 and K_2, and samples belonging to K_1 and K_2 span the two-dimensional linear subspaces V_1 and V_2, respectively, as shown in Fig. 1.

Now, the query $\boldsymbol{x}_q$ is newly measured, and its class should be identified as K_1 or K_2. The correlation coefficient can be used as the index of sample discrimination. For example, $\boldsymbol{x}_1 \in K_1$ and $\boldsymbol{x}_2 \in K_2$ are selected from each class in a random manner, the correlation coefficients $C_{1,q}$ and $C_{2,q}$ are calculated respectively, and the class including the sample with the largest absolute value of the correlation coefficient can be identified as the class of $\boldsymbol{x}_q$.

2.2. The Q and T^2 statistics

Although several indices of similarity between data sets have been proposed (Kano, Ohno, Hasebe, & Hashimoto, 2001; Kano, Hasebe, Hashimoto, & Ohno, 2002), the Q statistic is used as an index of sample discrimination in this work. The Q statistic is derived from principal component analysis (PCA), which is a tool for data compression and information extraction (Jackson and Mudholkar (1979)). PCA finds linear combinations of variables that describe major trends in data.

In PCA, the loading matrix $\boldsymbol{V}_R \in \Re^{M\times R}$ is derived as the right singular matrix of a data matrix $\boldsymbol{X} \in \Re^{N\times M}$ whose ith row is $\boldsymbol{x}_i^T$, and the column space of $\boldsymbol{V}_R$ is the subspace spanned by principal components. Here, M, N, and $R(\leq M)$ denote the numbers of variables, samples, and principal components retained in the PCA model, respectively. All variables are mean-centered and appropriately scaled. The score is a projection of $\boldsymbol{X}$ onto the subspace spanned by principal components. The score matrix $\boldsymbol{T}_R \in \Re^{N\times R}$ is given by

$$\boldsymbol{T}_R = \boldsymbol{X}\boldsymbol{V}_R. \tag{2}$$

$\boldsymbol{X}$ can be reconstructed or estimated from $\boldsymbol{T}_R$ with linear transformation $\boldsymbol{V}_R$:

$$\hat{\boldsymbol{X}} = \boldsymbol{T}_R\boldsymbol{V}_R^{\mathrm{T}} = \boldsymbol{X}\boldsymbol{V}_R\boldsymbol{V}_R^{\mathrm{T}}. \tag{3}$$

The information lost by the dimensional compression, that is, errors, is written as

$$\boldsymbol{E} = \boldsymbol{X} - \hat{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{V}_R\boldsymbol{V}_R^{\mathrm{T}}). \tag{4}$$

The Q statistic is defined as

$$Q = \sum_{m=1}^{M} (x_m - \hat{x}_m)^2 \tag{5}$$

where x_m is the m th variable of $\boldsymbol{x}$ and $\hat{x}_m$ is the estimate of x_m. The Q statistic is the distance between the sample and the subspace spanned by principal components. In other words, the Q statistic is a measure of dissimilarity between the sample and the modeling data from the viewpoint of the correlation among variables.

In addition, to guarantee that a sample is located in modeling data and to avoid extrapolation, Hotelling's T^2 statistic can be used:

$$T^2 = \sum_{r=1}^{R} \frac{t_r^2}{\sigma_{t_r}^2} \tag{6}$$

where σ_{t_r} denotes the standard deviation of the rth score t_r. The T^2 statistic expresses the normalized distance from the origin in the subspace spanned by principal components. The Q and T^2 statistics can be integrated into a single index (Raich, & Cinar, 1994):

$$J = \lambda T^2 + (1 - \lambda) Q \tag{7}$$

where $0 \leq \lambda \leq 1$.

3. Nearest Correlation Method

The NN method and the k-means method can discriminate or cluster samples on the basis of the distance without a teacher signal. However, they do not take into account the correlation among variables. In this section, a new unsupervised pattern recognition method based on the correlation

among variables, referred to as the nearest correlation (NC) method, is proposed. In the proposed NC method, sample geometry is used for sample discrimination.

3.1. Concept of the NC method

Suppose that the hyper-plane P in Fig. 2 (left) expresses the correlation among variables and the samples on P have the same correlation. Although samples $\boldsymbol{x}_1$ to $\boldsymbol{x}_5$ have the same correlation and they are on P, samples $\boldsymbol{x}_6$ and $\boldsymbol{x}_7$ have different correlation from the others. The NC method aims to detect samples whose correlation is similar to the newly measured query $\boldsymbol{x}_q$. In this example, $\boldsymbol{x}_1$ to $\boldsymbol{x}_5$ on P should be detected.

At first, the whole space is translated so that the query becomes the origin. That is, $\boldsymbol{x}_q$ is subtracted from all samples $\boldsymbol{x}_i (i = 1, 2, \cdots, 7)$. Since the hyper-plane P is translated to the plane containing the origin, it becomes the linear subspace V.

Next, a line connecting each sample and the origin is drawn. Suppose another sample can be found on this line. In this case, $\boldsymbol{x}_1$-$\boldsymbol{x}_4$ and $\boldsymbol{x}_2$-$\boldsymbol{x}_3$ satisfy such a relationship as shown in Fig. 2 (right). The correlation coefficients of these pairs of samples must be 1 or -1. On the other hand, $\boldsymbol{x}_6$ and $\boldsymbol{x}_7$ that are not the elements of V cannot make such pairs. Therefore, the samples of the pairs whose correlation coefficients are ± 1 are thought to have the same correlation as $\boldsymbol{x}_q$.

However, $\boldsymbol{x}_5$ that does not make a pair cannot be detected by this method even though it is on V. To detect $\boldsymbol{x}_5$, a linear subspace is derived from the selected pairs by using PCA, and the derived linear subspace corresponds to V.

Finally, the Q statistics for all samples $\boldsymbol{x}_i$ $(i = 1, 2, \cdots, 7)$ are calculated by using the PCA model expressing V. The samples with small Q statistics are located close to the linear subspace V, and such samples have correlation similar to the query. Although $\boldsymbol{x}_5$ cannot be detected in the previous step, it can be detected in this step because its Q statistic is 0. On the other hand, $\boldsymbol{x}_6$ and $\boldsymbol{x}_7$ are not detected in this step since they have large Q statistics.

In addition, the T^2 statistic can be used to take into account the distance from the origin. In the present work, J in Eq. (7) is used as the index for sample selection. The samples with small J are selected as the samples similar to the query.

In the implementation of the above procedure, the threshold of the correlation coefficient γ $(0 < \gamma \leq 1)$ has to be used since there are no pairs whose correlation coefficient is strictly ± 1. That is, the pairs should be selected when the absolute values of their correlation coefficients are larger than γ.

3.2. Algorithm of the NC method

Assume that the samples stored in the database are $\boldsymbol{x}_n \in \Re^M$ $(n = 1, 2, \cdots, N)$ and the query is $\boldsymbol{x}_q \in P$ $(\dim(P) = R)$. The samples belonging to P should be detected as similar samples to $\boldsymbol{x}_q$. The algorithm of the proposed NC method is as follows:

1. Set R, γ $(0 < \gamma \leq 1)$, δ $(0 < \delta)$, λ $(0 \leq \lambda \leq 1)$ and K or $\bar{J}$.
2. $\boldsymbol{x}'_n = \boldsymbol{x}_n - \boldsymbol{x}_q$ for $n = 1, 2, \cdots, N$.
3. Calculate the correlation coefficients $C_{k,l}$ between all possible pairs of $\boldsymbol{x}'_k$ and $\boldsymbol{x}'_l$ $(k \neq l)$.
4. Select the pairs satisfying $|C_{k,l}| \geq \gamma$, and set the number of the selected pairs S.

5. If $S < R$, then $\gamma = \gamma - \delta$ ($\delta > 0$) and return to step 4. If $S \geq R$, then go to the next step.
6. Arrange the samples of the pairs selected in step 4 as the rows of the matrix $\boldsymbol{X}'$.
7. Derive the linear subspace V from $\boldsymbol{X}'$ by using PCA. The number of principal components is R.
8. Calculate the index J of $\boldsymbol{x}'_n$, and $J_n = J$ for $n = 1, 2, \cdots, N$.
9. Detect the first K samples in ascending order of J_n or the samples whose J_n is smaller than $\bar{J}$ as samples similar to the query $\boldsymbol{x}_q$, where $\bar{J}$ is the threshold.

In step 5, when S is smaller than R, the threshold γ has to be relaxed to increase the number of selected pairs since the linear subspace V is not spanned by the samples of the selected pairs. R can be used as the tuning parameter. In addition, in step 1, the default values of parameters γ and δ can be determined as $1 - 10^{-3}$ and 10^{-3}, respectively.

3.3. Numerical example

The discrimination performance of the proposed NC method is compared with that of the NN method through a numerical example. In this example, data consist of three classes that have different correlations, and the samples belonging to the same class as the query should be detected. The discrimination rate is defined as

$$\text{Discrimination Rate } [\%] = \frac{L}{K} \times 100 \tag{8}$$

where K is the number of detected samples and L ($L \leq K$) is the number of samples that belong to the same class as the query among the detected

samples. Samples in each of three classes are generated by using the following equation.

$$\boldsymbol{x}_i = \boldsymbol{A}_i\boldsymbol{s} + \boldsymbol{n} \quad (i = 1, 2, 3) \tag{9}$$

$$\boldsymbol{s} = [s_1\ s_2]^T \tag{10}$$

$$\boldsymbol{n} = [n_1\ n_2\ n_3\ n_4\ n_5]^T \tag{11}$$

where $\boldsymbol{A}_i \in \Re^{5\times 2}$ is a coefficient matrix, $s_i \sim N(0, 10)$ and $n_i \sim N(0, 0.1)$. $N(m, \sigma)$ is the random number following the normal distribution whose mean is m and standard deviation is σ. The coefficient matrices are as follows:

$$\boldsymbol{A}_1 = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 1 \\ 2 & 3 \\ 1 & 3 \end{bmatrix},\ \boldsymbol{A}_2 = \begin{bmatrix} 3 & 3 \\ 2 & 1 \\ 3 & 1 \\ 3 & 2 \\ 2 & 0 \end{bmatrix},\ \boldsymbol{A}_3 = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 1 & 3 \\ 0 & 4 \\ 3 & 1 \end{bmatrix}. \tag{12}$$

100 samples are generated in each of three classes. In addition, a query belonging to each class is prepared. The number of detected samples K is fixed at 20.

In this example, $R = 2$, $\lambda = 0$, γ and σ are the default values. Sample generation and sample detection by the NN method and the NC method

Table 1: Discrimination performance of the NC method and the NN method

	Discrimination rate [%]			CPU time [ms]
	Class 1	Class 2	Class 3	
NC method	97.5	95.9	96.9	13.9
NN method	74.0	78.3	65.2	1.1

are repeated 100 times and the average discrimination rates [%], and the average CPU time [ms] are calculated. The computer configuration used in this numerical example is as follows: OS: Windows Vista Business (64bit), CPU: Intel Core2 Duo 6300 (1.86GHz×2), RAM: 2G byte, and MATLAB® 7.5.0 (2008a).

Table 1 shows the discrimination results of the NN method and the NC method. The proposed NC method can achieve higher discrimination performance than the NN method. On the other hand, the computational load of the NC method is relatively heavy since singular value decomposition (SVD) is used for calculating the correlation among variables. In fact, the computation of SVD occupies most of the computation time of the NC method.

4. Correlation-based Just-In-Time Modeling

The conventional JIT modeling uses the distance for sample selection when a temporary local model is constructed. However, its estimation performance is not always high since it does not take into account the correlation among variables. Recently, the Correlation-based JIT (CoJIT) modeling that selects samples for local modeling on the basis of the correlation among variables has been proposed(Fujiwara et al., 2009).

Figure 3 shows the difference of sample selection for local modeling between JIT modeling and CoJIT modeling. The samples are classified into two groups that have different correlations. In the conventional JIT modeling, samples are selected regardless of the difference of correlation as shown in Fig. 3 (left), since a neighbor region around the query is defined only by distance. On the other hand, CoJIT modeling can select samples whose

correlation is best fit for the query as shown in Fig. 3 (right).

The procedure of CoJIT modeling is as follows: 1) several data sets are generated from data stored in the database. 2) The index J is calculated from the query and each data set. 3) The data set whose J is the smallest is selected. 4) A temporary local model is constructed from the selected data set.

In the above procedure, each data set is generated so that it consists of successive samples included in a certain period of time using moving time-window, because the correlation in such a data set is expected to be very similar (Fujiwara et al., 2009). However, CoJIT modeling is applicable to only time-series data due to its data set construction.

On the other hand, the NC method can detect samples that have correlation similar to the query regardless of whether the object data is time-series data or not. This is the motivation for integrating the proposed NC method with CoJIT modeling. Using the NC method for constructing data set for local modeling, CoJIT modeling is applicable to also other than time-series data. This integrated soft-sensor design method is referred to as NC-CoJIT modeling.

Assume that the sampling interval of the output is longer than that of the input, and the output at time t, $\boldsymbol{y}_t$, should be estimated. Now, the input and the output measured at the same time are stored in the database, and the sth input-output sample $\boldsymbol{x}^{\{s\}} \in \Re^M$ $(s = 1, 2, \cdots, S)$ and $\boldsymbol{y}^{\{s\}} \in \Re^L$ are stored as matrices $\boldsymbol{X}_S \in \Re^{S \times M}$ and $\boldsymbol{Y}_S \in \Re^{S \times L}$, respectively. To cope with process dynamics, measurements at different sampling times can be included in $\boldsymbol{x}^{\{s\}}$. The algorithm of the proposed NC-CoJIT modeling is as follows:

1. When the input at time t, $\boldsymbol{x}_t$, is measured, the index J is calculated from $\boldsymbol{x}_t$ and $\boldsymbol{X}_{t-1}$ that was used for building the previous local model f_{t-1}, and $J_I = J$.
2. If $J_I \leq \bar{J}_I$, $f_t = f_{t-1}$, $\boldsymbol{X}_t = \boldsymbol{X}_{t-1}$, and f_t is used for estimating the output $\boldsymbol{y}_t$. Then, return to step 1. If $J_I > \bar{J}_I$, go to the next step. Here, $\bar{J}_I$ is the threshold.
3. K input samples whose correlation is similar to the query are detected from $\boldsymbol{X}_S$ by the NC method, and they are arranged as the rows of $\boldsymbol{X}_t \in \Re^{K \times M}$. In addition, K output samples corresponding to the detected input samples are selected from $\boldsymbol{Y}_S$, and they are arranged as the rows of $\boldsymbol{Y}_t \in \Re^{K \times L}$, where K is the number of the detected samples.
4. A new local model f_t whose input is $\boldsymbol{X}_t$ and output is $\boldsymbol{Y}_t$ is built.
5. The output $\boldsymbol{y}_t$ is estimated by using f_t.
6. The above steps 1 through 5 are repeated until the next output sample $\boldsymbol{y}_{S+1}$ is measured. When $\boldsymbol{y}_{S+1}$ is measured, $\boldsymbol{y}_{S+1}$ and its corresponding input $\boldsymbol{x}_{S+1}$ are stored in the database, and return to step 1.

In the above algorithm, any modeling method can be used for building a local model f. In the present work, partial least squares regression (PLS) is used to cope with the colinearity problem. In addition, steps 1 and 2 control the model update frequency. When the threshold $\bar{J}_I$ is large, the update frequency becomes low. The local model is updated every time when new input measurements are available in the case where $\bar{J}_I = 0$.

5. Case Study

In this section, the estimation performance of the proposed NC-CoJIT modeling is compared with that of the conventional JIT modeling through their applications to product composition estimation for a parallelized chemical reaction process. The detailed model used in this case study is described in Johannesmeyer, & Seborg (1999).

5.1. Problem setting

In this process, two reactors R1 and R2 are operated in parallel. Although these reactors have the same structure as shown in Fig. 4, they have different characteristics. In each reactor, an irreversible reaction A $\longrightarrow$ B takes place. The set point of the reactor temperature $T^{[d]}(d = 1, 2)$ is independently changed between $\pm$2K every ten days. Although 15 process variables listed in Table 2 are calculated in the simulations, measurements of only five variables $T^{[d]}$, $h^{[d]}$, $Q^{[d]}$, $Q_C^{[d]}$, $Q_F^{[d]}$ are used for analysis, and their sampling interval is one minute. In addition, reactant concentration $C_A^{[d]}$ is measured in a laboratory once a day.

To take into account catalyst deactivation and fouling as changes in process characteristics and individual difference of each reactor, the frequency factor $k_0^{[d]}$ and the heat transfer coefficient $U^{[d]}$ are assumed to decrease with time. In addition, $k_0^{[d]}$ and $U^{[d]}$ are recovered every half year (180 days). Figure 5 shows changes of the frequency factors $k_0^{[d]}$ and the heat transfer coefficients $U^{[d]}$. The operation data of each reactor for a half year (180 days) were stored in the database.

In this case study, a soft-sensor for estimating reactant concentration of

Table 2: Process variables of the chemical reaction process

Variable	Caption
C_A	Reactant concentration [mol/m^3]
T	Reactor temperature [K]
T_C	Coolant temperature [K]
h	Reactor level [m]
Q	Reactor exit flow rate [m^3/min]
Q_C	Coolant flow rate [m^3/min]
Q_F	Reactor feed flow rate [m^3/min]
C_{AF}	Feed concentration [mol/m^3]
T_F	Feed temperature [K]
T_{CF}	Coolant feed temperature [K]
hC	Level controller instruction
QC	Outlet flow rate controller instruction
TC	Temperature controller instruction
QC_C	Colorant flow rate controller instruction
T_{set}	Reactor temperature set point [K]

the newly developed reactor R3 is designed. The estimation starts at the 90th day after the start of its operation, and the soft-sensor is updated in the next half year. Although R3 has only a small amount of data due to its short operation term, the soft-sensor is updated by searching samples similar to the current operation of R3 from the other reactor operation data in the past.

5.2. Estimation result

The reactant concentration $C_A^{[3]}$ is estimated by JIT modeling and the proposed NC-CoJIT modeling.

In JIT modeling, a local model is constructed from samples located in a neighbor region around the query, whenever the input variables are measured. A new sample is stored only when $C_A^{[3]}$ is measured. In this case study, linear local models are built and Euclidean distance is used as the measure for selecting samples to build local models. To take into account process dynamics, the input data consist of the present sample and the sample measured one minute before. The MATLAB® Lazy Learning Toolbox developed by Bontempi, Birattari, & Bersini (1999b) is used.

The estimation result of JIT modeling is shown in Fig. 6. The top figure shows the estimation result for 180 days. Although $C_A^{[3]}$ is estimated every minute, only estimates corresponding to the measurements of $C_A^{[3]}$ are plotted to compare the estimates with the measurements. The bottom figure shows the enlarged result for two months before and after the catalyst recovery. The estimates shown in the bottom figure are calculated every minute whenever the input variables are observed, and they fluctuate by measurement noise. In this figure, r denotes the correlation coefficient between measurements and estimates, and RMSE is the root-mean-squared error.

This result shows that JIT modeling cannot achieve high estimation performance. The reason for the poor performance of JIT modeling seems to be that it does not take account of correlation among variables when a local model is built. To validate this reasoning, NC-CoJIT modeling is applied to the same problem.

In NC-CoJIT modeling, samples for local modeling are selected by the NC method, and PLS are used for model building. The parameters of the NC method are determined by trial and error, $R = 6$, $\lambda = 0.01$, γ and σ are the default values, and the parameter for update frequency is $\bar{J}_I = 0$. To take into account process dynamics, the input data consist of the present sample and the sample measured one minute before.

The estimation result of NC-CoJIT modeling, in Fig. 7, clearly shows that the estimation performance of the proposed NC-CoJIT modeling is very high. RMSE is improved by about 35% in comparison with JIT modeling. Actually, the number of the selected samples from each reactor in the first day of the estimation period is as follows: the number of samples from R1, R2 and R3 is 14, 19 and 17, respectively. Although R3 has only 90 samples at the estimation start, the soft-sensor is constructed by using not only the samples from R3 but also similar samples from R1 and R2.

These results of this case study clearly show that the proposed NC-CoJIT modeling can cope with not only abrupt changes in process characteristics but also the individual difference of production devices. In addition, it can construct a high performance soft-sensor for a newly developed device, even when only a small amount of operation data is available.

6. Conclusion

A new unsupervised pattern recognition method that can detect samples whose correlation is similar to the query is proposed. In addition, CoJIT modeling is integrated with the proposed NC method. The proposed NC-CoJIT modeling can cope with not only changes in process characteristics

but also the individual difference of production devices and improve the estimation performance since it can select samples for local modeling by appropriately accounting for the correlation among variables. The proposed NC-CoJIT modeling has the potential for realizing efficient maintenance of soft-sensors.

References

Amirthalingam, R., & Lee, J. (1999). Subspace Identification Based Inferential Control Applied to a Continuous Pulp Digester. J Process Control, *9*(5), 397-406.

Atkeson, CG., Moore, AW., & Schaal, S. (1997). Locally Weighted Learning, Artif Intell Rev, *11*(1-5), 11-73.

Bontempi, G., Birattari, M., & Bersini, H. (1999a). Lazy Learning for Local Modeling and Control Design. Int J Control, *72*(7-8), 643-658.

Bontempi, G., Birattari, M., & Bersini, H. (1999b). Lazy Learners at Work: The Lazy Learning Toolbox. EUFIT'99, Aachen, Germany. Sep.13-16.

Cheng, C., & Chiu, MS. (2004). A New Data-Based Methodology for Nonlinear Process Modeling. Chem Eng Sci, *59*(13), 2801-2810.

Desai, K., Badhe, Y., Tambe, SS., & Kulkarni, RD. (2006). Soft-sensor development for fed-batch bioreactors using support vector regression. Biochem Eng J, *27*(3), 225-239.

Fujiwara, K., Kano, M., & Hasebe, S. (2009). Soft-Sensor Development Using Correlation-Based Just-In-Time Modeling. AIChE J, *55*(7), 1754-1765.

Jackson, JE., & Mudholkar, GS. (1979). Control Procedures for Residuals Associated with Principal Component Analysis. Technometrics, *21*(3), 341-349.

Jain, P., Rahman, I., & Kulkarni, BD. (2007). Development of a soft sensor for a batch distillation column using support vector regression techniques. Chem Eng Res Des, *85*(A2), 283-287.

Johannesmeyer, M., & Seborg, DE. (1999). Abnormal Situation Analysis Using Pattern Recognition Techniques and Histrical Data. AIChE Annual meeting, Dallas, TX, Oct.31-Nov.5.

Kamohara, H., Takinami. A., Takeda. M., Kano, M., Hasebe, S., & Hashimoto, I. (2004). Product Quality Estimation and Operating Condition Monitoring for Industrial Ethylene Fractionator. J Chem Eng Jpn, *37*(3), 422-428.

Kano, M., Hasebe, S., Hashimoto, I., & Ohno H. (2002). Statistical Process Monitoring Based on Dissimilarity of Process Data. AIChE J, *48*(6), 1231-1240.

Kano, M., Lee, S., & Hasebe, S. (2009). Two-Stage Subspace Identification for Softsensor Design and Disturbance Estimation. J Process Control, *19*(2), 179-186.

Kano, M., Miyazaki, K., Hasebe, S., & Hashimoto, I. (2000). Inferential Control System of Distillation Compositions Using Dynamic Partial Least Squares Regression. J Process Control, *10*(2-3), 157-166.

Kano, M., & Nakagawa, Y,. (2008). Data-Based Process Monitoring, Process Control, and Quality Improvement. Recent Developments and Applications in Steel Industry, Comput Chem Eng, *32*(1-2), 12-24.

Kano, M., & Ogawa, M., (2009). The State of the Art in Advanced Chemical Process Control in Japan. ADCHEM (CD-ROM), Istanbul, Turkey, July 12-15.

Kano, M., Ohno, H., Hasebe, S., & Hashimoto, I. (2001). A New Multivariate Statistical Process Monitoring Method using Principal Component Analysis. Comput Chem Eng, *25*(7-8), 1103-1113.

Khatibisepehr .S, & Huang, B. (2008). Dealing with irregular data in soft sensors: Bayesian method and comparative study. Ind Eng Chem Res, *47*(22), 8713-8723.

Kohonen, T. (2001). Self-organizing maps. New York, Springer, 3rd edition.

Kresta, VJ., Marlin, TE., & MacGregor, JF. (1994). Development of Inferential Process Models Using PLS. Comput Chem Eng, *18*(7), 597-611.

Mejdell, T., & Skogestad. S. (1991). Estimation of Distillation Compositions from Multiple Temperature Measurements Using Partial-Least-Squares Regression. Ind Eng Chem Res, *30*(12), 2543-2555.

Qin, SJ. (1998) . Recursive PLS Algorithms for Adaptive Data Modeling. Comput Chem Eng, *22*(4-5), 503-514.

Radhakrishnan, V., & Mohamed, A. (2000). Neural networks for the Identifi-

cation and Control of Blast Furnace Hot Metal Quality. J Process Control, *10*(6), 509-524.

Raich, A., & Cinar, A. (1994). Statistical Process Monitoring and Disturbance Diagnosis in Multivariable Continuous Processes. AIChE J, *42*(4), 995-1009.

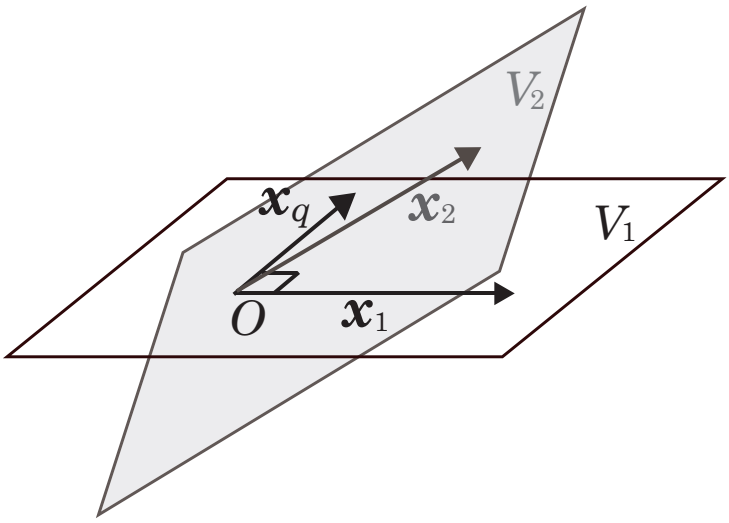

Figure 1: An example of vector geometry in 3-dimensional space

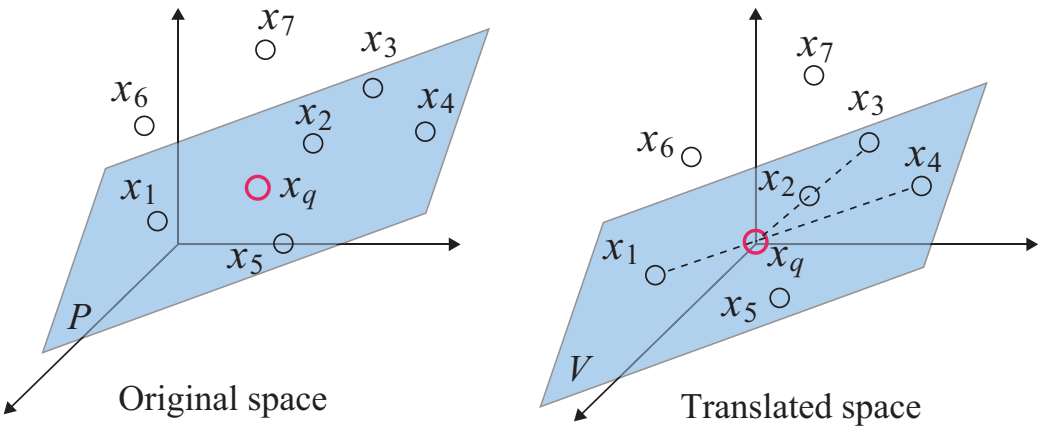

Figure 2: An example of the procedure of the NC method

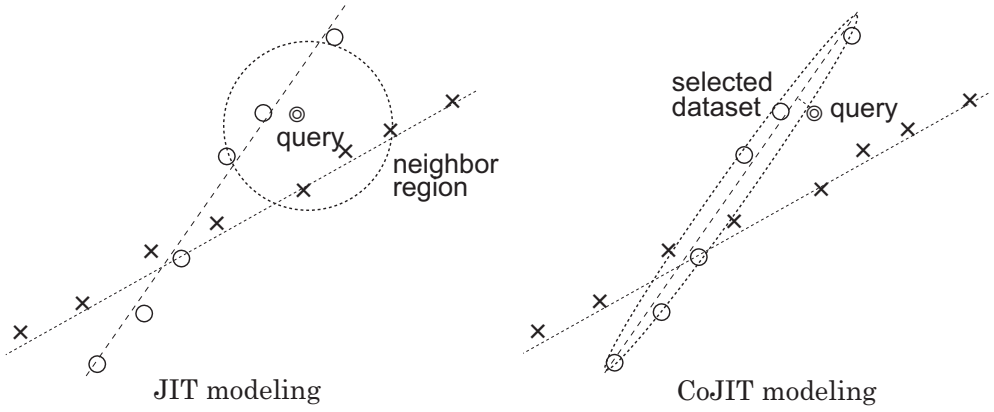

Figure 3: Sample selection in JIT modeling (left) and CoJIT modeling (right)

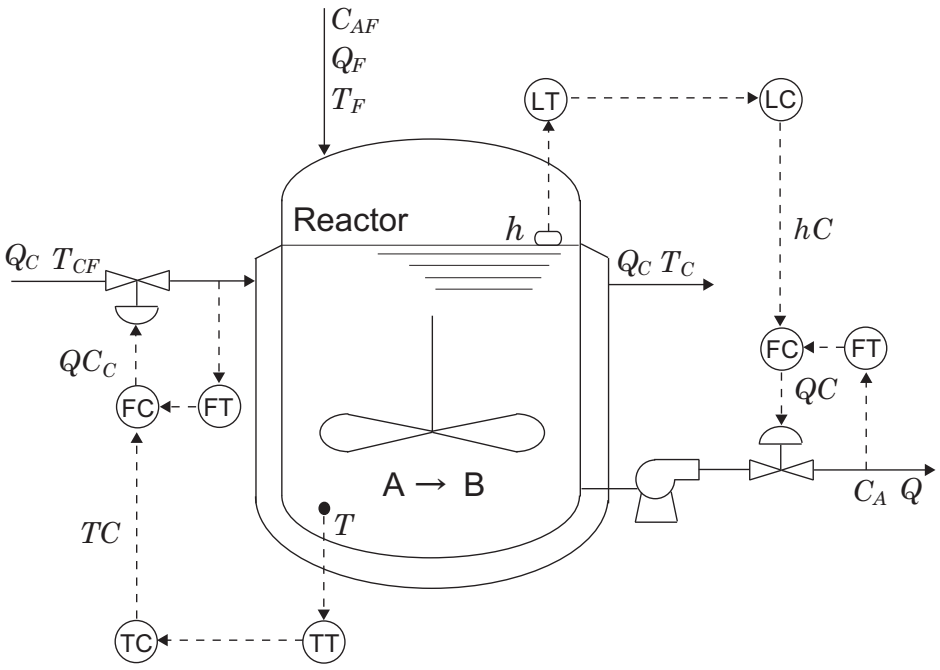

Figure 4: Schematic diagram of the chemical reaction process with cascade control systems

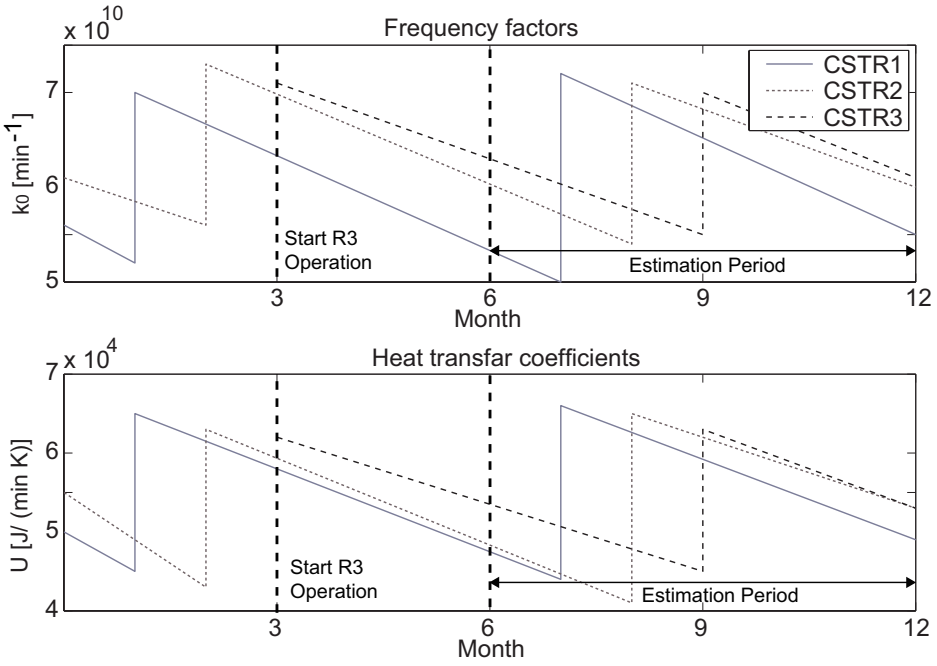

Figure 5: Changes of frequency factors and heat transfer coefficients of the chemical reaction processes

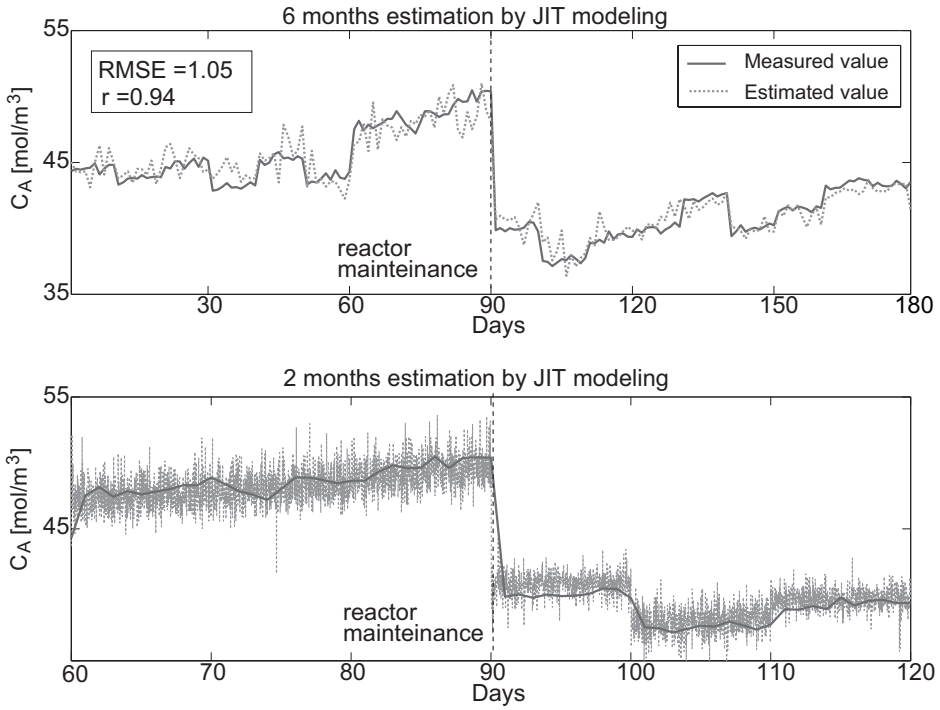

Figure 6: Prediction result of $C_A^{[3]}$ by JIT modeling

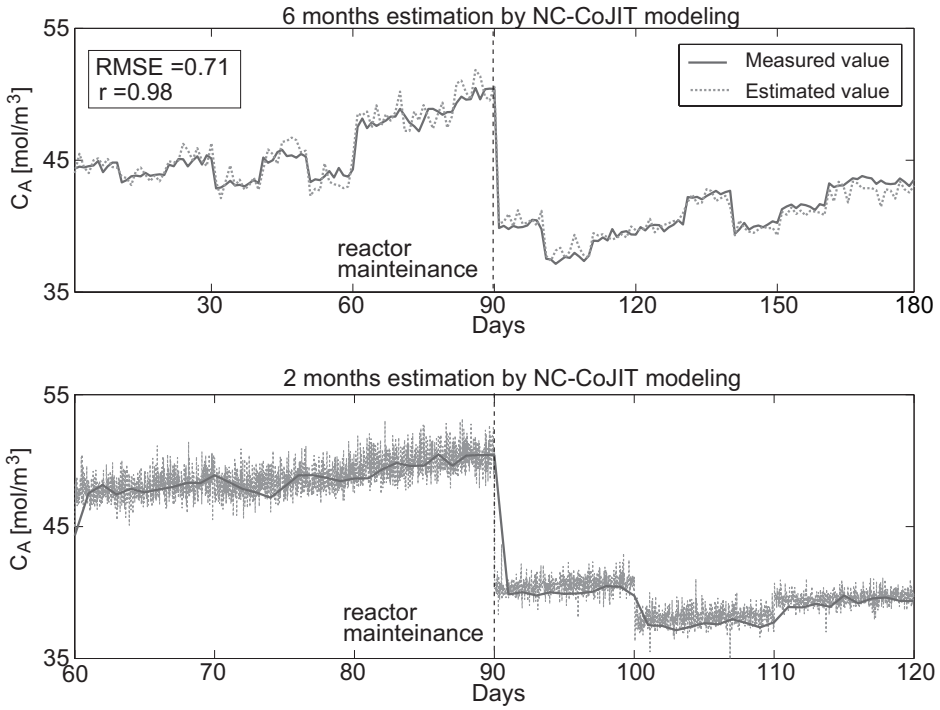

Figure 7: Prediction result of $C_A^{[3]}$ by NC-CoJIT modeling