[생물화공심포지움]

Data-driven modeling of the anaerobic wastewater treatment plant using robust adaptive dynamic PLS method

(2004년 7월 26일(월) 16:00~16:40)

박 종 문

(포항공과대학교)

Data-driven modeling of the anaerobic wastewater treatment plant using robust adaptive dynamic PLS method

Hae Woo Lee[a], Min Woo Lee[b], Jea Youl Joung[a], Jong Moon Park[a]*

[a]Department of Chemical Engineering, [b]School of Environmental Engineering,
POSTECH, San 31 Hyoja Dong, Pohang 790-784, South Korea

요 약

Principal Component Analysis나 Partial Least Squares와 같은 다변량 통계 기법은 변수간의 correlation structure로부터 공정의 variance를 설명할 수 있는 latent variable를 얻고 이를 이용하여 공정을 효과적으로 modeling할 수 있는 방법으로 최근 들어 많은 관심을 얻고 있다. 하지만 PLS는 공정이 stationary state에 있다고 가정하기 때문에, 생물학적 공정의 non-stationary and time-varying behavior를 설명하기에 부적절하다. 본 논문에서는 PLS 알고리즘의 혐기성 폐수처리 공정에의 적용에 있어, 이와 같은 문제를 해결하기 위해서 adaptive PLS 알고리즘을 사용함으로써 변화하는 공정의 특성에 대응하여 모델을 update하는 방법을 이용하였다. 하지만 실시간 데이터로부터 adaptive PLS 방법을 적용하는 데에는 많은 어려움이 존재하며, 특히 outlier나 abnormal disturbance에 모델이 부적절하게 adaptation하는 문제가 발생할 수 있다. 따라서 이의 해결을 위해 adaptive PLS를 적용하는데 있어 robustness를 향상시키기 위해 monitoring index를 이용하여 abnormal data에 weight를 주고 안정적인 모델의 update가 가능하게 하는 방법을 제안하였으며, 이를 적용하여 성공적으로 혐기성 폐수처리 공정의 output을 예측하고 효과적으로 공정을 모니터링할 수 있었다. 만들어진 PLS 모델은 산업폐수를 처리하기 위한 industrial plant에서 측정된 실제 데이터에 적용하여 그 효용성을 입증하였으며, 그 결과는 mechanistic model을 적용하기 힘든 실공정에 비교적 쉽게 implementation할 수 있는 장점이 있다.

1. 서 론

혐기성 폐수 처리공정은 혐기 조건에서 미생물의 복잡한 일련의 제거 기작에 의해 폐수내의 유기물이 biogas로 변환되는 과정을 거치며, 다른 전통적인 호기 폐수처리 방법에 비해 여러 장점을 가지기 때문에, 최근 많은 관심을 얻고 있다. 특히, 고농도의 폐수처리능력을 지니고 있으며, 난분해성 물질의 처리에 용이하고, 슬러지의 생산이 적으며, 에너지가 적게 들고, 생산된 biogas내에 포함된 메탄가스를 에너지원으로 이용할 수 있다는 데 장점이 있다.[Bernard et al, 2001] 하지만, 혐기성 처리공정의 운전에 있어 methanogen의 생장속도가 매우 느리고 독성물질이나 과부하 등에 매우 불안정한 특성을 지니고 있기 때문에, 혐기성 공정의 안정적인 제어와 모니터링이 매우 중요하며 이에 대한 많은 연구가 이루어지고 있다.[Steyer et al, 1999]

따라서, 효과적으로 혐기성 공정을 모니터링하기 위해서는 신뢰할 수 있는 dynamic model의 개발이 필요하다. 이를 위해 많은 연구자들이 혐기성 처리 시스템에 대한 모델을 연구하였으며, 그 결과 IAW 그룹은 최근의 연구결과를 종합하여 anaerobic digestion model을 발표하였다. [IWA Task Group, 2002] 하지만 생물학적 공정의 경우, 반응 mechanism 자체가 매우 복잡한 미생물 기작과 물리화학적인 상호작용으로 이루어지기 때문에 물리화학적 지식과 물질수지식에 바탕을 둔 mechanistic model을 실제로 적용하기가 매우 어렵거나, 복잡한 구조와 많은 수의 파라미터들로 인해 이를 이용하는데 많은 시간과 노력이 필요하다. 특히 혐기성 공정에서 많이 쓰이고 있는 fixed bed나 fludized bed의 경우에 복잡한 hydrodynamics나 inhibition 반응들을 수식적으로 표현하기 위해서는 아직도 많은 연구가 필요한 실정이다. [Lubbert et al, 2001]

최근 들어 센서와 컴퓨터 기술의 발달로 대부분의 공정에서는 많은 데이터들이 수집되고 저장되고 있으며, 이들을 적절히 이용함으로써 공정을 쉽게 파악하고 효율적으로 제어하기 위한 많은 연구들이 이루어지고 있다. 따라서 공정에서 측정되는 input-output 데이터의 관계로부터 효과적으로 필요한 정보를 추출해내고 이들간의 상호관계와 통계학적인 방법론을 응용하는 multivariate latent variable modeling을 이용할 경우 보다 효과적으로 대상 시스템을 분석할 수 있다. 이러한 다변량 통계분석에서 많이 쓰이는 대표적인 기법으로는 PCA, PLS 등이 있으며, 이들을 이용한 성공적인 공정 모니터링[Wise, 1996; MacGregor, 1995; Kourti, 1995], 공정 모델링 및 indentification[Shi, 2000; Wise, 1990?; Dayal, 1996; Juricek, 2002]에 대한 결과들이 보고되고 있다.

최근에는 주로 회분식 공정을 중심으로 하여 MPCA를 이용한 다변량 공정 모니터링에 대한 연구논문들이 많이 발표되고 있으며[Lee, 2002], 이를 산업 현장에 응용하려는 시도들이 긍정적으로 평가받고 있다. 하지만 이들 다변량 통계분석을 혐기성 공정을 비롯한 생물공정에 응용함에 있어 어려운 문제는 미생물의 활성과 직접적으로 관련된 state variable들을 측정하기가 힘들고, 폐수처리 시설이 기본적으로 가지는 특성으로 인해 influent의 조건을 일정하기 유지하기 힘들며, 다른 여타 공정들에 비해 disturbance가 크고 운전환경에 따라 미생물의 활성이 변화하여 공정이 시간에 따라

변화하는 특성을 가진다는 것 등이다. 따라서 이러한 문제들을 해결하기 위해서는 혐기성 공정에서 일반적으로 측정되는 여러 physical quantity와 관련된 측정변수들을 효과적으로 이용하여 control, monitoring에 적용하여야 하며, 이를 위해 기본적으로 필요한 dynamic model은 시간에 따른 process change에 대한 영향을 고려할 수 있어야 한다는 점이다.

pervision을 위해 neural network [tay et al, 2000; Lee, 2000], pattern recognition [Marsili-Libelli, 1996], fuzzy-logic [steyer, 2001] 등이 함께 이용되기도 한다.

하지만 이러한 conventional PCA/PLS 기법은 공정이 stationary state에 있다라는 가정에 기초하고 있기 때문에, 이들 방법을 실공정에 적용하는 데 있어 큰 걸림돌이 되고 있다. 대부분의 화학공정은 시간에 따라 공정이 변화하는 non-stationary property를 가지고 있으며, 특히 생물학적 폐수처리공정의 경우, 유입수의 유량이나 농도의 disturbance가 심하고 non-stationary하며, 운전조건에 따라 미생물이 환경에 적응하여 변화하기 때문에 이에 대한 고려없이 다변량 통계기법을 적용할 경우, 잘못된 결과를 불러올 수 있다.[Lennox, 2002] 따라서 time-varying and non-stationary problem을 해결하기 위해 시간에 따라 모델을 update하여 새로운 process state를 모델에 포함시키는 방법이 필요하다.

이러한 문제를 해결하기 위해 exponentially weighed moving average (EMWA) PCA/PLS [Wold, 1994], recursive PCA [Li, 2000], recursive PLS [Helland, 1991; Dayal, 1997] 등이 개발되었으며, Qin et al. (1998)에서는 기존의 방법을 개선하여 block 단위의 데이터에 moving window와 forgetting factor를 적용하여 모델을 update하고 computational load를 줄일 수 있는 방법을 제안하였다. 이 알고리즘은 다양한 non-stationary and time varying process에 적용될 수 있으며, 각 변수들의 mean, variance, correlation structure 변화에 적응하여 공정의 모니터링[Wang, 2003] 뿐만 아니라 quality prediction, process identification 및 process control [Vijaysai , 2003] 등에도 이용될 수 있다. 따라서 최근 생물학적 폐수처리 공정에도 시간에 따른 공정의 변화를 고려한 adaptive PCA/PLS 기법들이 적용되고 있으며 [Rosen, 2001; Lennox, 2002; Lee, 2002], 경우에 따라서는 서로 다른 time scale에서 일어나는 process dynamics를 capture하기 위해 multiresolution analysis(MRA)를 함께 이용하기도 한다. [Rosen, 2001; Teppola, 2000]
하지만 실제로 이러한 adaptive method를 실공정에 온라인으로 적용하기 위해서는 몇가지 고려해야 할 사항들이 있으며, Wang et al.(2003)은 adaptive method를 적용함에 있어 생각할 수 있는 문제점들을 다음과 같이 지적하였다.

a) the process monitoring results can be different for different window size
b) a heavy computational effort may be required for updating the model
c) previous data that is representative of normal process behavior is discounted in favor of new data which may not be as representative

위의 문제 이외에도, 실시간으로 새로운 데이터가 축적이 됨에 따라 모델을 update할 경우 사용하게 되는 data의 quality corruption의 문제도 생각할 수 있다. Window size의 경우, window size가 커질수록 모델이 안정화되지만 공정의 변화에 재빨리 적응하지 못하게 되고, window size가 작아지면 그 반대의 문제가 생기게 되므로 trade-off problem으로 볼 수 있다. Computational load를 줄이기 위해서는 일정한 시간 간격의 block 단위로 모델을 update하거나 [Qin, 1998], improved PLS 알고리즘을 이용하는 것이 가능하다. [Dayal, 1997]. 또한 discount of old data의 경우, variable forgetting factor를 적용하거나 [Dayal, 1997], 모델을 update할 때 사용하는 새로운 data의 quality를 유지함으로써 해결할 수 있다. 이때, 모델에 사용하는 data의 quality를 유지하는 문제는 모델의 stability와도 연관이 있으며, 실시간으로 알고리즘을 적용할 경우, normal state의 데이터와 abnormal state의 데이터를 구분하여 normal state의 데이터만을 선택적으로 모델의 update에 사용하여야 reasonable한 결과를 얻을 수 있다. Dayal et al. (1997)의 결과를 보면, recursive PLS를 이용하여 output data를 예측할 때, 불충분한 정보가 들어있는 새로운 데이터를 사용함으로써, 몇몇 구간에서 모델의 정확성이 떨어지는 blowups현상을 관찰할 수 있었다. 이러한 결과는 실시간으로 데이터가 축적될 경우, 생길 수 있는 sensor fault나 unknown disturbance 등에 의해 더욱 심각해질 수 있다. 따라서 모델을 update하는 과정에서 이러한 문제를 해결하고 모델의 stability를 높이기 위해서는 robust PLS algorithm이 적용되어야 한다.

본 논문에서는 anaerobic wastewater treatment process의 quality prediction과 모니터링을 위해 adaptive moving window PLS method를 적용하여 보았다. 먼저 장기간의 데이터를 사용하여 process change에 의한 conventional PLS 알고리즘의 영향을 살펴보고 이를 해결하기 위해 Qin et al.(1998)이 제안한 recursive PLS 알고리즘을 이용하였다. 그 후, 실시간으로 이 방법을 적용할 경우, 생길 수 있는 모델의 blowups를 방지하고 새롭게 모델에 포함되는 데이터의 quality를 유지하기 위한 modified robust adaptive PLS algorithm을 제안하였다. 이때, 사용한 데이터는 PTA 제조공정에서 생산되는 폐수를 처리하기 위한 industrial anaerobic wastewater treatment plant에서 실시간 acquisition system에 의해 측정된 값들을 이용하였다.

2. 이 론

Partial Least Square (PLS)

공정에서 측정되는 데이터를 input, $X \in \Re^{m \times n}$와 output, $Y \in \Re^{m \times p}$로 두면, 이들간의 관계가 linear하다고 가정할 때 다음과 같이 나타낼 수 있다.

$$Y = XC + V$$

이때 C는 X와 Y간의 관계를 나타내는 regression coefficient이며, V는 noise를 나타낸다. 이때, 변수간에 강한 상관관계가 존재할 경우, least squared solution, $C=(X^TX)^{-1}X^TY$ 는 ill-conditioned되게 된다. 이 문제를 해결하기 위해 PLS 알고리즘은 X와 Y를 다음과 같이 low dimensional space로 project하여, 새로운 orthogonal latent variable, t_a, u_a 를 얻는다.

$$X=\sum_{a=1}^{A} t_a p_a^T + E = TP^T + E = \hat{X} + E \tag{2}$$

$$Y=\sum_{a=1}^{A} u_a q_a^T + F = UQ^T + F = \hat{Y} + F \tag{3}$$

이때 얻어진 벡터 t_a, u_a 를 각각 y-score vector, u-score vector라고 하며, 벡터 p_a, q_a 를 loading vector라고 부른다. PLS알고리즘은 predictor block의 variation을 최대화하면서predicted block과의 correlation도 고려하여 X와 Y의 covariance를 최대화하는 방향으로 loading과 score를 구하게 된다. 위의 outer relation에 의해 X와 Y는 rank one matrices의 sum of series로 나타내어지게 되며, 이때 구해지는 loading의 개수 A가 min(m,n)과 같게 되면 E와 F는 zero가 되면서 ordinary least square로 수렴된다. 하지만 대부분의 화학공정에서는 데이터의 variation이 몇몇 common cause variance에 의해 좌우되기 때문에 A를 min(m,n)보다 작게 함으로써 noise와 collinearity를 줄일 수 있게 된다.

PLS알고리즘은 다음과 같은 inner relation에 의해 predicted u-score vector, $\hat{u}_a$ 를 t-score vector, t_a 로부터 추정하게 된다.

$$\hat{u}_a = b_a t_a \tag{4}$$

이를 matrix form으로 나타내면 다음과 같은 관계에 의해 output, Y를 예측할 수 있게 된다.

$$Y=\sum_{a=1}^{A} \hat{u}_a q_a + F = \sum_{a=1}^{A} b_a t_a q_a + F = TBQ^T + F \tag{5}$$

이때, B는 scalar, b_a 를 diagonal entry로 갖고 off-diagonal element는 0인 $n \times n$ matrix이다.

PLS알고리즘에서 score와 loading vector를 구하기 해 많은 알고리즘들이 제안되었으며, 그 중에서 가장 많이 쓰이는 것으로는 NIPALS [Geradi et al, 1986]이 있고, 그 외에도 SIMPLS [de Jong, 1993], kernel algorithm [Dayal, 1997] 등이 쓰인다. Qin et al. (1998)은 recursive PLS알고리즘을 좀 더 쉽고 빠르게 계산하기 위해 NIPALS 알고리즘에 수정을 가해 loading, p_a과 weight, w_a을 normalize 하는 대신 t-score를 normalize하는 방법을 이용하였으며, 그 과정은 Table 1과 같다.

Model에 필요한 factor의 수를 구하는 방법은 여러 가지가 있지만 [Jackson, 1991] 보통 cross validation [Wold, 1978]을 통해 구하는 것이 가장 일반적이다. Cross validation에서는 데이터를 g개의 그룹으로 나눈 후, (g-1) block의 데이터로부터 모델을 만들고 이때 제외된 block의 데이터를 사용하여 모델을 test하고 PRESS(predictive error sum of squares) 값을 계산한다. 이 과정을 각각의 block에 대해 g번 반복하여 total PRESS를 구하고, 다시 이 과정을 각각의 PC 수에 대해 반복하여 series of PRESS value를 얻는다. 이때 주어진 PRESS 값으로부터 optimal number of PC를 얻기 위해서는 minimum PRESS 값에 해당하는 PC수를 사용하는 방법이 많이 쓰이지만, 이 방법을 실공정으로부터 얻어진 데이터에 사용할 경우 noise나 faulty condition등으로 인해 잘못된 결과를 가져올 수 있게 된다. [Osten, 1988] 따라서 본 논문에서는 fist local minimum of PRESS 값에 해당하는 PC 수까지르 사용하는 Wold's R criteria [Li, 2002]를 stopping rule로 이용하였다.

Table 1. Modified NIPALS algorithm proposed by Qin

1. Scale X and Y to zero-mean and unit-variance.
2. Initialize $E_0 = X, F_0 = Y$ and $h = 0$.
3. Let $h = h + 1$ and take u_k as some column of F_{h-1}.
4. Iterate the PLS outer model until it converges

$$w_h = E_{h-1}^T u_h / u_h^T u_h$$

$$t_h = E_{h-1} w_h / \| E_{h-1} w_h \|$$

$$q_h = F_{h-1}^T t_h / \| F_{h-1}^T t_h \|$$

$$u_h = F_{h-1} q_h$$

5. Calculate the X-loadings : $p_h = E_{h-1}^T t_h$
6. Find inner model : $b_h = u_h^T t_h$
7. calculate the residuals : $E_h = E_{h-1} - t_h p_h^T, F_h = F_{h-1} - b_h t_h q_h^T$
8. Return to step 2 until all principal factors are calculated

PLS 알고리즘은 각 변수간의 상관관계를 고려하여 X와 Y의 covariance를 최대화한다. 공정에서 측정되는 변수들은 mass balance나 fundamental physicochemical principle 등에 의해 각 변수간에 큰 상관관계가 존재할 뿐만 아니라 closed-loop control structure, process dynamics, time-scale dependent event 등에 의해 변수 자체 내에 serial correlation 이 존재하기도 한다.

따라서 변수 내의 serial correlation을 고려하고 dynamic operation condition을 모델링하기 위해서는 다음과 matrix augmentation에 의해 dynamic model을 구성하여야 한다. [Ku, 1995]

$$X(t) = [x^T(t-1), x^T(t-2), \ \ldots \ , x^T(t-n_x)]$$
$$Y(t) = [y^T(t)] \tag{6}$$

이때, $x(t)$ 와 $y(t)$는 각각 time, t에서의 input, output vector를 나타내며, n_x는 dynamic nature of process를 나타내기 위한 time-lag order이다. 위의 관계에 의해 구해진 input-output relation은 FIR(finite impulse response) model에 의해 다음과 같이 나타낼 수 있다.

$$y(t) = a_1 x(t-1) + a_2 x(t-2) + \quad \ldots \quad + a_{n_x} x(t-n_x) + \varepsilon(t) = \sum_{i=1}^{n_x} a_i x(t-i) + \varepsilon(t) \tag{7}$$

X block에 output을 다음과 같이 augmentation할 경우, ARX(autoregressive with exogenous inputs) model의 identification 또한 가능하다. [Wise, 1990]

$$X(t) = [y^T(t-1), \ \ldots \ , y^T(t-n_y) \ | x^T(t-1), \ \ldots \ , x^T(t-n_x)] \tag{8}$$

이때, 얻어지는 ARX 모델의 input-output relation은 다음과 같이 나타낼 수 있다.

$$y(t) = \sum_{i=1}^{n_x} a_i x(t-i) + \sum_{j=1}^{n_y} b_j y(t-j) + \varepsilon(t) \tag{9}$$

Block-wise adaptive moving window PLS

Conventional PLS는 공정이 stationary state에 있다라고 가정하고 모델을 만들 때 사용한 데이터로부터 얻어진 parameter들을 이용하여, 새로운 데이터, X로부터 Y값을 예측하고 t-score value를 구하게 된다. 하지만 폐수처리 공정을 포함한 대부분의 화학공정은 시간에 따라 공정의 상태가 변화하고, disturbance, change of operation strategies 등에 의해 non-stationary state에 있는 경우가 대부분이다. Qin et al. (1998)은 Helland et al. (1992)에 의해 처음으로 제안된 recursive PLS 알고리즘을 개선하여 공정의 변화를 기존의 PLS 모델에 포함시켜 모델을 update하는

알고리즘을 다음과 같이 제안하였다.

먼저 공정으로부터 process variable과 quality variable로 구성된 데이터 행렬 $\{X_{old}, Y_{old}\}$를 얻었다고 가정하면, PLS 알고리즘을 통해 다음과 같이 parameter들을 계산할 수 있게 된다.

$$\{X_{old}, Y_{old}\} \xrightarrow{PLS} \{T_{old}, W_{old}, P_{old}, B_{old}, Q_{old}\} \tag{10}$$

이때, T, W, P, Q는 각각 PLS의 score, weight, loading을 나타내며, B는 inner relation coefficient를 diagonal element로 갖는 square matrix이다. 이때, latent variable의 수가 X의 rank와 같다면 다음과 같은 관계가 성립한다.

$$\begin{aligned} XX^T &= PT^TTP^T = PP^T \\ X^TY &= P^TTBQ^T + PT^TF = PBQ^T \end{aligned} \tag{11}$$

그후, 새로운 데이터 block, $\{X_{new}, Y_{new}\}$이 들어오면 다음과 같이 matrix를 구성하여 새로 PLS 모델을 update할 수 있다.

$$X = \begin{bmatrix} X_{old} \\ X_{new} \end{bmatrix}; \quad Y = \begin{bmatrix} Y_{old} \\ Y_{new} \end{bmatrix} \tag{12}$$

이때, $Y = XC + F$의 관계로부터 C는 다음과 같이 구할 수 있다.

$$\begin{aligned} C_{new}^{PLS} &= \left(\begin{bmatrix} X_{old} \\ X_{new} \end{bmatrix}^T \begin{bmatrix} X_{old} \\ X_{new} \end{bmatrix} \right)^+ \begin{bmatrix} X_{old} \\ X_{new} \end{bmatrix}^T \begin{bmatrix} Y_{old} \\ Y_{new} \end{bmatrix} \\ &= (X_{old}^TX_{old} + X_{new}^TX_{new})^+ (X_{old}^TY_{old} + X_{new}^TY_{new}) \end{aligned} \tag{13}$$

따라서 $\{X_{new}, Y_{new}\} \xrightarrow{PLS} \{T_{new}, W_{new}, P_{new}, B_{new}, Q_{new}\}$의 관계로부터 새로운 데이터 block에 대한 local PLS parameter를 얻는다면, 식 (11)에 의해 식 (13)는 다음과 같이 오직 PLS parameter matrix들로부터 계산할 수 있게 된다.

$$C_{new}^{PLS} = (P_{old}P_{old}^T + P_{new}P_{new}^T)^+ (P_{old}B_{old}Q_{old}^T + P_{new}B_{new}Q_{new}^T)$$
$$= \left(\begin{bmatrix} P_{old}^T \\ P_{new}^T \end{bmatrix}^T \begin{bmatrix} P_{old}^T \\ P_{new}^T \end{bmatrix} \right)^+ \begin{bmatrix} P_{old}^T \\ P_{new}^T \end{bmatrix}^T \begin{bmatrix} B_{old}Q_{old}^T \\ B_{new}Q_{new}^T \end{bmatrix} \quad (14)$$

따라서 recursive PLS 알고리즘에서는 PLS 모델을 update하기 위해 old data와 new data를 사용하는 대신 다음과 같은 matrix를 사용한다. 이때 중요한 것은 recursive PLS를 행할 때, maximum possible number of factor을 사용해야 한다는 점이다.

$$\left\{ \begin{bmatrix} P_{old}^T \\ P_{new}^T \end{bmatrix}, \begin{bmatrix} B_{old}Q_{old}^T \\ B_{new}Q_{new}^T \end{bmatrix} \right\} \xrightarrow{PLS} \{T, W, P, B, Q\} \quad (15)$$

위의 알고리즘을 이용하면 conventional PLS을 통해 모델을 update하는 것에 비해 PLS run-size를 줄일 수 있으므로 모델을 update하기 위한 computational load를 줄일 수 있으며, memory size 또한 절약할 수 있다.

본 논문에서는 위의 알고리즘을 이용하여 Fig 1. 과 같은 adaptive moving window PLS를 사용하였다. $\{X_{new}, Y_{new}\}$의 block size를 1로 할 경우, 위의 알고리즘은 recursive form을 지니게 된다. 하지만 이 경우, Wold et al. (1994)에서 지적한 바와 같이, matrix size가 작아지고, 모델이 자주 update될수록, 새로 들어오는 데이터로 인해 loading이 rotation하여 모델 structure가 destabilized 될 가능성이 커지게 된다. 실제로 block size를 1로 잡고 recursive PLS를 수행하여 본 결과, noise와 faulty condition 등으로 인해 모델이 update되면서 점점 destabilized 되는 현상을 직접 관찰할 수 있었다. 따라서 PLS 모델이 안정적인 결과를 내기 위해서는 block size를 충분히 크게 잡아주어 불필요한 모델의 update로 인해 모델이 rotation되면서 불안정해지는 것을 막아주어야 한다.

또한 moving window를 적용할 경우, window size에 따라 서로 다른 결과를 얻을 수 있기 때문에 적절히 process dynamic를 나타낼 수 있도록 그 크기를 정해주어야 한다. 일반적으로 window size가 커질수록 sample의 수가 커져서 모델이 stable하게 되고, variance가 줄어들게 되지만, process change에 대한 적응이 느려지는 delay를 가지게 된다. 반대로 window size가 작아지면 process change에 대한 적응은 빨라지지만, quality가 일정하지 않은 데이터가 들어올 경우, 모델의 구조가 불안정해지고, 이전 데이터에 존재하고 있던 process dynamics에 대한 중요한 정보를 빨리 잃어버려서 잘못된 결과를 얻을 수 있다. 따라서 적절한 window size의 문제는 adaptive moving window PLS의 적용에 있어 중요하게 고려해야 할 사항이다.

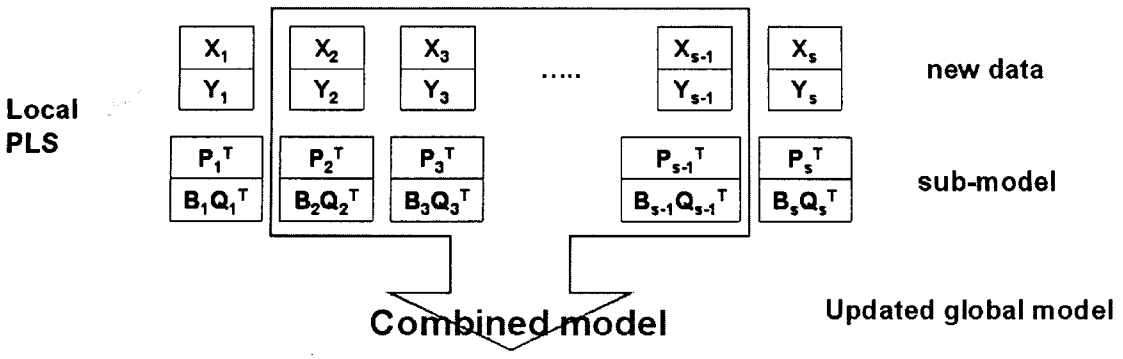

Fig 1. schemes for adaptive PLS algorithm

Table 2. the adaptive moving window PLS algorithm

1. Formulate the data matrix $\{X, Y\}$. Autoscale the data
2. Derive a PLS model using the algorithm in Table 1 : $\{X, Y\} \xrightarrow{PLS} \{T, W, P, B, Q\}$
3. When a new pair of data, $\{X_n, Y_n\}$ is available, scale it the same ways it was done in step 1. Perform PLS to derive a submodel :

   $$\{X_n, Y_n\} \xrightarrow{PLS} \{T_n, W_n, P_n, B_n, Q_n\}$$

4. Formulate $X = \begin{bmatrix} P_i^T \\ P_{i+1}^T \\ \cdots \\ P_n^T \end{bmatrix}$, $Y = \begin{bmatrix} B_i Q_i^T \\ B_{i+1} Q_{i+1}^T \\ \cdots \\ B_n Q_n^T \end{bmatrix}$, and return to step 2. (Discard previous block continuously to maintain the window size)

Monitoring statistics for PLS

PLS 모델을 사용하여 얻은 loading과 score vector를 사용하면 PCA와 동일한 알고리즘을 통해 공정을 모니터링하고 fault를 진단하는 것이 가능하다. Normal state의 공정 데이터를 사용하여 PLS 모델을 정의한 후, 새로운 데이터를 주어진 PLS 모델에 projection시켜 얻은 score value를 reference set의 값과 비교하여 공정의 상태를 진단할 수 있다. 이때 현 시점에서의 공정의 상태가 정상상태에서 얼마만큼 deviate했는지 감지하기 위해 다음의 두 index를 사용한다.

$$T^2 = \sum_{i=1}^{A} \frac{t_i^2}{\lambda_i} = x^T P \Lambda^{-1} P^T x \tag{16}$$

$$SPE = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 = \left\| (I - PP^T)x \right\|^2 = \left\| \tilde{C}x \right\|^2 \tag{17}$$

이때, λ_i는 i번째 score vector, t_i의 variance를 나타내며, Λ는 PLS 모델에서 구한 A번째 score vector까지의 λ_i 값을 diagonal element로 하는 matrix이다. T^2는 PLS 모델이 정의하는 PCS(principal component space)에서의 deviation 정도를 나타내며, SPE는 PLS 모델이 정의하는 PCS가 나타내지 못한 RS(residual space)에서의 deviation 정도를 나타낸다.

따라서 위의 두 index를 complementary하게 이용하여 모니터링 차트를 구성함으로써 공정의 variance 정도를 나타낼 수 있다. 이때 T^2과 SPE의 confidence limit, χ_l^2과 δ^2는 각각의 parametric distribution을 이용하여 구할 수 있다. [Jackson, 1991] 본 논문에서는 공정의 변화에 따라 confidence limit 또한 adaptation하도록 하기 위해 Wang et al.(2003)이 제안한 방법을 이용하여 다음과 같이 confidence limit를 계산하였다.

$$\varsigma^{1-\alpha} = g\chi^2(\alpha, h) \tag{18}$$

이때 $\varsigma^{1-\alpha}$는 구하고자 하는 monitoring statistics의 confidence limit를 나타내며, χ^2는 confidence, $(1-\alpha)$에 대한 Chi-Squared distribution function을 나타낸다. 또한 g와 h는 과거의 monitoring index 값이 포함한 moving window로부터 얻어진 mean value, $\bar{\varsigma}$과 variance, σ_ς^2로부터 다음과 같이 구할 수 있다.

$$g = \frac{\sigma_\varsigma^2}{2\bar{\varsigma}}, \quad h = \frac{2\bar{\varsigma}^2}{\sigma_\varsigma^2} \tag{19}$$

monitoring index에 대한 adaptive confidence limit를 구함으로써 공정의 변화에 의한 영향을 모델에 포함시키는 것이 가능하며, Wang et al.(2003)은 위와 같은 방법을 통해 구한 confidence limit를 사용하여 false alarm을 줄이고 공정의 non-stationary and time-varying behavior를 효과적으로 고려할 수 있음을 보여주었다.

3. Robustness issue of PLS algorithm

PLS 모델을 만들 때 사용하는 데이터에 포함되는 outlier는 process disturbance나 sensor fault 또는 erroneous setting of operation strategy 등으로 인해 발생하며, 이러한 outlier는 모델의 예측성능에 크게 영향을 미친다. 특히 실시간으로 모델을 update할 경우, 정상적인 상태의 data와 outlier를 구분하는 것은 매우 중요하며, 이 과정이 적적하게 수행되지 않을 경우, 모델을 통해 얻은

결과의 잘못된 해석으로 인해 적절한 공정의 감시와 예측을 이룰 수 없게 된다. 하지만 공정 데이터의 multi-collinear한 특성과 multiple outlier의 존재로 인해 이러한 outlier를 구분하는 것은 쉽지 않은 일이며, 이를 위해 robust PLS 알고리즘을 이용하는 연구들이 많이 이루어지고 있다.

wakeling (1992), Griep (1995)은 NIPALS 알고리즘 내의 regression 과정에 직접 weighted regression을 사용하는 robust PLS 알고리즘을 제안하였으며, Commins et al. (1995)은 IRLS(iteratively reweighted least squares)의 방법에서 출발하여 sample에 직접 weight를 주는 iteratively reweighted partial least squares method를 제안하였다. 이때 weighting function은 regression residual의 함수로 표현되며 효율이나 robustification의 정도에 따라 여러 가지 형태의 weighting function을 사용할 수 있다. Wakeling et al(1992)에서는 다음과 같은 weighting function을 사용하여 결과가 수렴할 때까지 계속 iteration을 행함으로써 robust PLS 모델을 구하는 방법을 제안하였다.

$$\omega_i = [1-(r_i / kf\%)^2]^2 \quad for \quad |r_i| < kf\%$$

$$\omega_i = 0 \quad for \quad |r_i| > kf\% \tag{20}$$

이때 ω_i는 sample weight를 나타내며, r_i는 regression residual을, $f\%$는 median of absolute values of the residuals을 나타낸다. 또한 k는 sensitivity factor로써 6에서 9까지의 값을 쓰는 것이 보통이다.

본 논문에서는 commins et al(1995)에서 제안한 IRPLS 알고리즘을 modify하여 sample에 weight를 주는 방법을 이용함으로써 실시간으로 update되는 모델의 robustness를 높이고자 하였다.

4. Application to anaerobic wastewater treatment plants

Process description

본 논문에서는 adaptive moving window PLS를 사용하여 industrial scale의 혐기성 폐수처리시설의 output quality를 예측하고 performance를 모니터링하여 보았다. 혐기성 폐수처리 시설은 xylene을 원료로 사용하여 PTA를 생산하는 석유화학공장에서 생산되는 다량의 aromatic acid를 포함한 폐수를 처리하고 있으며, 자세한 공정도는 Fig 2.와 같다.

폐수의 주요 성분으로는 product로 생산되는 terephthalic acid와 반응 부산물로 생성되는 benzoic acid, p-toluic acid, acetic acid, orthophthlic acid, trimellitic acid 등이 있으며, 평균 유입 TOD 농도는 약 8500 ppm 정도로 운전되고 있다. 공정에서 발생된 폐수는 Fig 2. 와 같이 settling

tank를 거친 후, down-flow type의 anaerobic treatment filter reactor로 유입된다. 이 과정에서 반응기로 유입되는 폐수의 농도와 pH를 조절하기 위해 별도의 고농도 폐수와 NaOH가 투여되며, 혐기성 반응기는 honey-comb type의 담체를 이용한 biofilm system으로 구성되어 있다. 반응기를 거쳐 처리된 유출수의 일부는 반응기 내의 flow distribution을 원활하게 하고, 유입부하의 disturbance를 완충하기 위한 목적으로 반응기로 반송되며, 나머지는 후단부에 위치한 활성오니 공정을 거친 후, 최종 방출된다.

공정 중에 측정되는 모든 변수들은 실시간 data acquisition system에 의해 data base로 저장되며, 공정의 운전은 이들 정보를 이용하여 현장 operator들에 의해 제어되고 있다. 따라서 효율적인 공정의 monitoring과 quality의 예측은 operator들이 disturbance나 sensor fault 등의 상황에 대해 재빠르게 대처하기 위한 의사결정을 support하는데 있어 매우 중요한 요소로 작용할 수 있다.

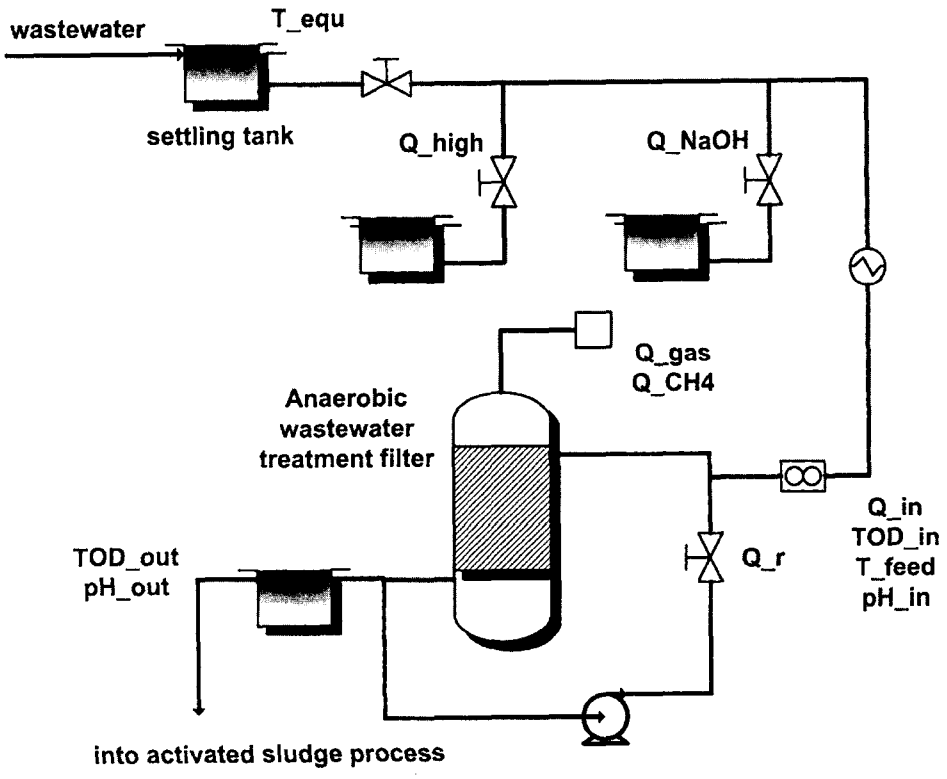

Fig 2. schematic diagram of anaerobic wastewater treatment plant

Modeling approach

Adaptive PLS modeling에 사용된 데이터는 2003년 7월부터 2004년 2월까지 현장에서 측정되었으며, sampling interval은 1시간으로 총 4369개의 observation으로 이루어져있다. PLS

모델을 만들기 위한 변수로는 Table 3과 같이 predictor variable, X로 15개, predicted dependent variable, Y로 2개의 변수를 사용하였다. Predictor variable는 현장에서 실시간으로 측정되는 10개의 변수와 이들의 nonlinear relationship으로부터 계산된 5개의 추가변수로 구성되어 있으며, predicted variable로는 주요공정 지표로 활용되는 유출수의 TOD 농도와 에너지원으로 재사용되는 메탄가스의 발생량으로 이루어져있다. 위의 변수들을 사용하여 PLS 모델을 만들기 위해서는 변수들의 Sampling interval와 HRT (약 50여 시간)를 감안하고 변수들 간에 autocorrelation과 dynamic condition을 모델에 포함시키기 위해 식 (7)이나 (9)를 사용하여 dynamic model을 구성하여야 한다. 따라서 본 논문에서는 predictor variable의 과거치를 matrix에 augmentation하는 FIR type의 모델을 구성하였다. ARX 모델의 경우, FIR 모델에 비해 더욱 robust하고 parsimonious한 parameter를 얻을 수 있다는 장점이 있지만, 모델의 structure가 공정의 dynamics를 잘 반영하지 못할 경우, Qin et al.(1998)에서 지적한 바와 같이 auto-regression term이 강조되는 경향이 존재하였으므로 이를 제외하였다. 하지만 predictor 변수의 항목에서 공정의 output에 해당하는 pH_out과 같은 변수를 사용하였고, rTOD_in과 같은 변수를 계산하기 위해서 TOD_out의 과거값을 사용하였기 때문에, 본 논문에서 이용한 모델은 엄격히 정의하면 semi-FIR 모델이라고 할 수 있을 것이다.

Table 3. variables used in the modeling of anaerobic wastewater treatment process

(* variables were calculated from measured variable using external knowledge of the process)

notation		description
Predictor Variable (X)	Q_in	Inflow rate to the reactor *(Ton/h r)*
	TOD_in	TOD measured at inlet *(mg/L)*
	pH_in	pH measured at inlet
	pH_out	pH measured at outlet
	T_equ	Temperature of equalization tank
	T_feed	Temperature of feed flow
	Q_r	Recycle flowrate *(Ton/hr)*
	Q_high	Inflow rate from concentrated feed tank to inlet *(Ton/hr)*
	Q_gas	Flowrate of biogas produced in the reactor *(m3/hr)*
	Q_NaOH	Feedrate of NaOH solution *(Ton/hr)*
	rTOD_in*	Actual TOD inflow rate *(mg/L) = (Q_in*TOD_in+Q_r*TOD_out)/(Q_in +Q_r)*
	TOD_load*	TOD loading rate *(g TOD/hr) = Q_in * TOD_in*
	CT*	Contact time *(hr) = volume of reactor / (Q_in+Q_r)*
	HRT*	Hydraulic retention time *(hr) = volume of reactor / Q_in*
	rTOD_load*	Actual TOD loading rate *(g TOD/hr) = rTOD_in * (Q_in+Q_r)*
Predicted variable (Y)	TOD_out	TOD measured at outlet *(mg/L)*
	Q_CH4	Flowrate of methane gas produced in the reactor *(m3/hr)*

FIR 모델을 구성하기 위해서는 적절한 time lag order의 값을 추정하는 것이 필요하다. Akaike's information method (AIC)는 이와 같은 상황에서 model order를 구하기 위해 사용할 수 있는 적절한 criteria를 제공하며, 다음과 같이 model의 goodness of fit과 complexity에 대한 penalty term으로 이루어져 있다.

$$AICc = n\log\sigma_\varepsilon^2 + 2m(n+1)/(m-n-2) \tag{21}$$

위의 식에서 m은 observation의 수, n은 변수의 수를 나타내며, σ_ε^2은 residual variance를 나타낸다. 일반적으로 AICc (bias-corrected AIC)는 small sample의 data set에 적합하며, large sample data의 경우는 BIC (Akaike's Baysian modification of AIC)가 더 우수한 결과를 낸다고 알려져 있다. [Wu, 1998]

$$BIC = n\log\sigma_\varepsilon^2 + (n-m)\log(1-n/m) + n\log m + n\log\{n^{-1}(\sigma_Y^2/\sigma_\varepsilon^2 - 1)\} \tag{22}$$

위의 식에서 σ_Y^2는 predicted variable, Y의 sample variance를 나타낸다. PLS 모델을 만들기 위해 사용한 데이터의 경우, sample size가 변수의 수에 비해 매우 크므로 BIC를 사용하여 모델의 time lag order를 추정하는 것이 적당하며, 그 결과는 Fig 3과 같다.

Fig 3의 결과를 보면 각각의 predicted variable에 대한 BIC의 변화경향은 거의 비슷하나, TOD_out은 time lag order가 8일 때, Q_CH4는 time lag order이 2일 때 최소값을 가지는 것을 볼 수 있다. 하지만 두 변수 모두, lag가 2일 때 first local minimum의 값을 가지므로, 보다 simple한 모델을 만들기 위해서 모델의 time lag order는 2로 결정하였다.

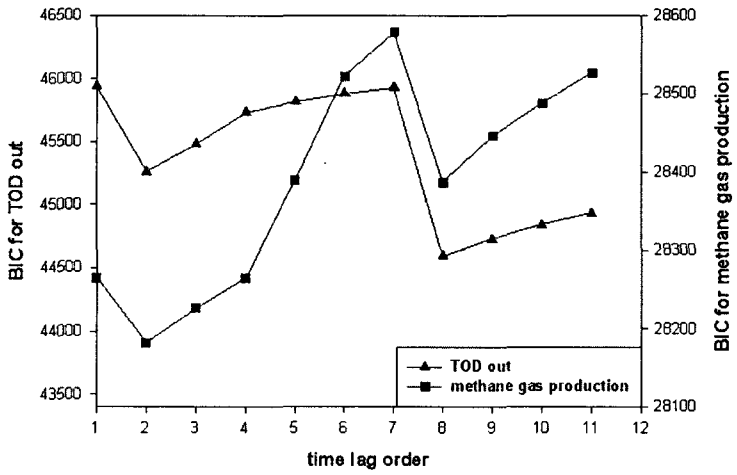

Fig 3. BIC value for various time lag order using whole data set and PLS algorithm

5. 결과 및 고찰

Conventional PLS algorithm

Adaptive moving window PLS 모델을 적용하기 위해 우선 1000개의 sample data를 사용하여 initial PLS 모델을 training시켰다. PLS 모델에서 optimal latent variable의 수는 cross-validation을 이용하여 3개를 사용하였으며 adaptive PLS 알고리즘과 conventional PLS 알고리즘을 이용할 때의 차이를 알아보기 위해, 이때 구한 PLS 모델을 나머지 validation set에 적용하여 output 변수를 예측하여 보았다. 많은 경우에 있어 FIR이나 ARX 모델을 identification하기 위해서 input signal에 PRBS(pseudo-random binary signal)을 주는 것이 보통이지만 industrial wastewater treatment plant의 경우 유입되는 폐수의 disturbance를 조정하기 힘들며, 혐기성 공정과 같이 민감한 시스템의 경우 input 조건을 마음대로 바꾸는 것이 어렵기 때문에, 그대신 충분한 size의 데이터를 사용하여 PLS 모델을 identification하여 주었다. 그럼에도 불구하고 Fig 4를 살펴보면 약 2300시간 후부터 모델의 예측 능력이 급격히 감소하며 bias를 보이는 것을 관찰할 수 있다. 따라서 이 시간대를 전후로 process의 변화가 일어났음을 추정할 수 있으며, 이러한 시간에 따른 공정의 상태변화가 모델에 포함되지 않을 경우, 공정의 모니터링이나 control에 잘못된 결과를 야기할 수 있음을 확인할 수 있다. 특히 미생물이 포함된 혐기성 폐수처리공정의 경우, 시간에 따라 미생물의 상태가 변화하거나 유입수의 조건이 전단의 생산공정에 크게 영향을 받기 때문에 기본적으로 공정이 시간에 따라 변화하는 특성을 지니고 있으며, stationary state에 머무르는 경우가 거의 없으므로 일정한 주기를 두고 모델을 update하는 것이 필수적이다.

위의 경우 이러한 공정의 변화 원인은 다변량 통계분석기법에서 자주 사용하는 monitoring chart나 contribution plot을 통해 쉽게 확인할 수 있으며, 이를 통해 확인된 몇몇 주요 변수들의 time profile은 Fig 5와 같다. Fig 5를 보면 변수들의 변화가 2300시간 대를 전후로 확연히 다른 패턴을 보이는 것을 볼 수 있다. 이러한 변화가 여러 변수들에서 확연히 드러나고 또한 일정시간 이상 지속되는 것으로 보아 단순한 disturbance가 아닌 operator에 의한 의도적인 operation strategy의 변화로 추정된다. 이러한 공정의 변화는 Fig 6의 PLS monitoring chart를 통해서도 확인이 가능하며, 공정의 변화가 SPE plot에서 두드러지게 나타나는 것으로 보아 변수들 간에 새로운 correlation이 생긴 것으로 해석할 수 있다.

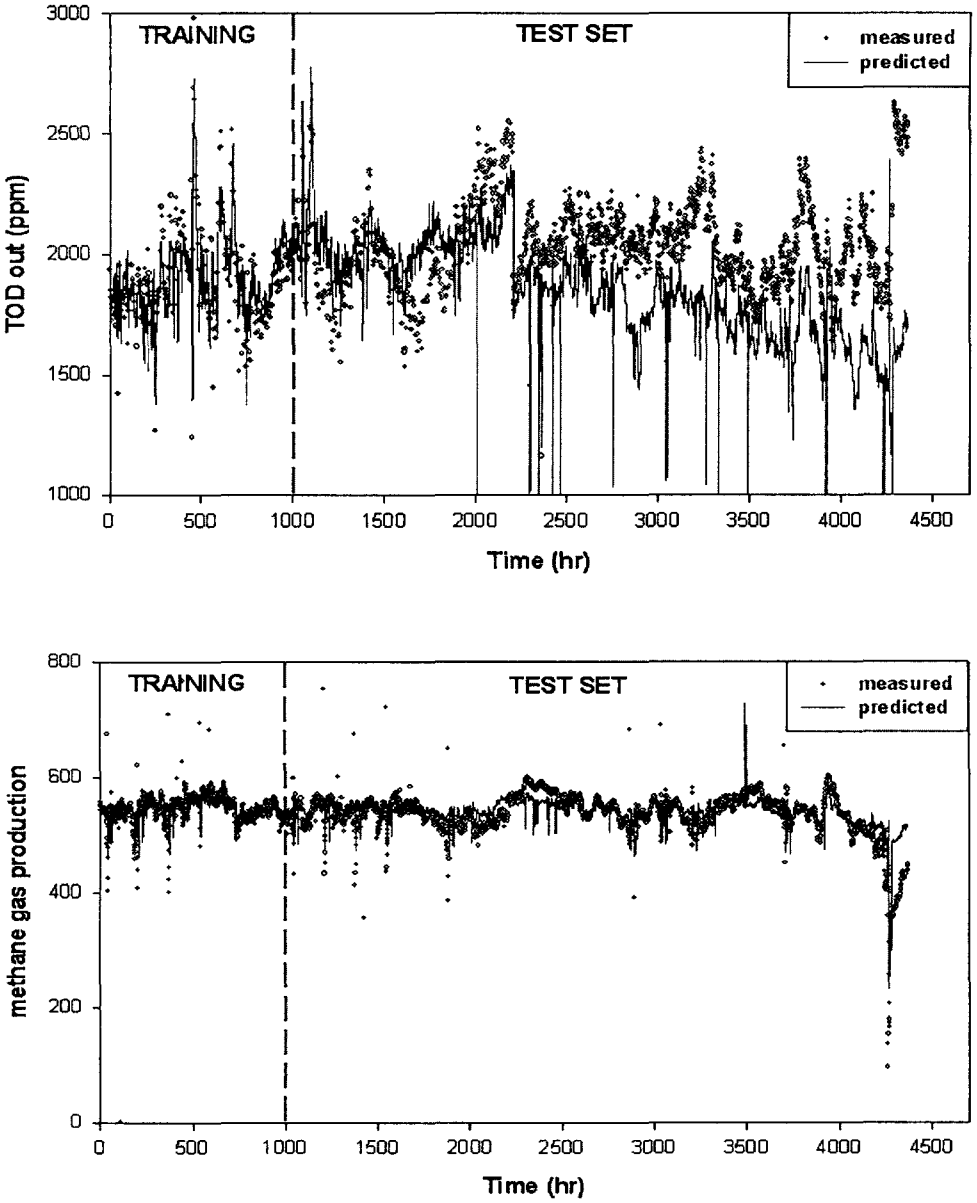

Fig 4. Prediction result using conventional PLS trained with first 1000 data sets

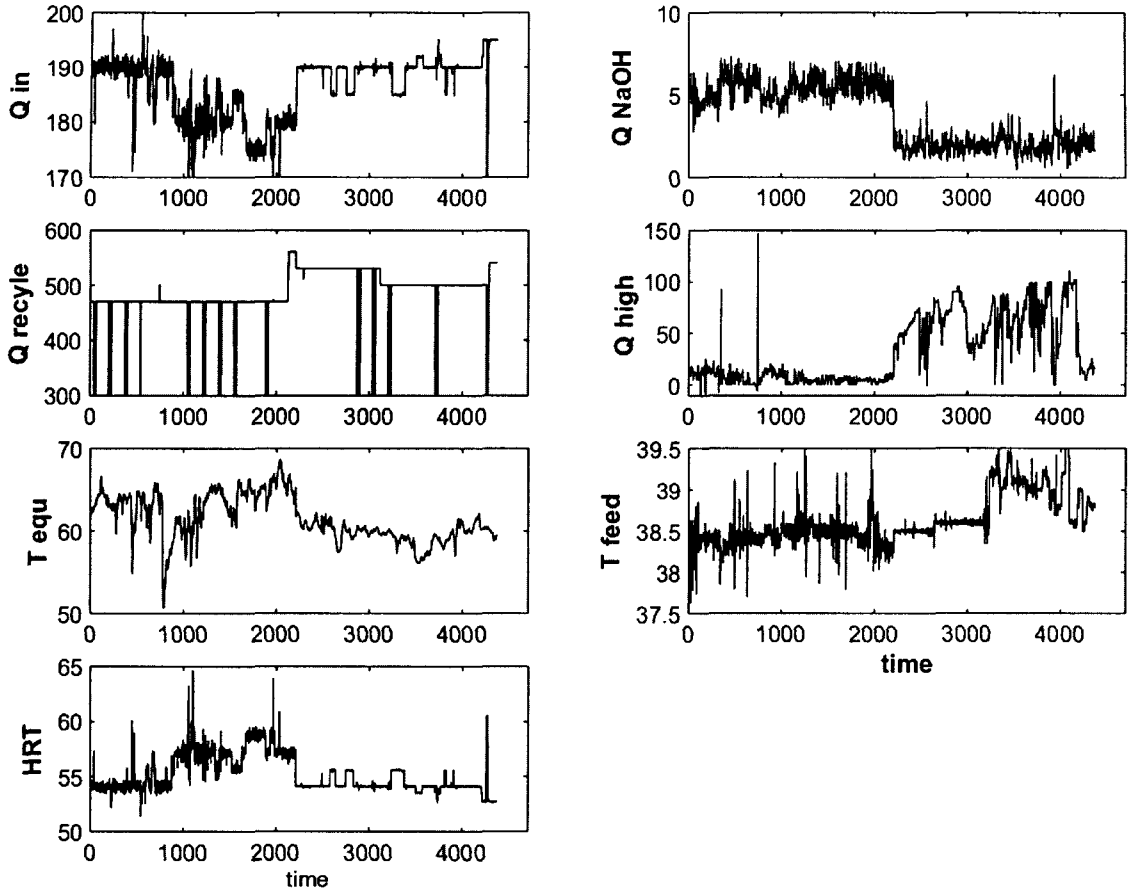

Fig 5. Time profile for various predictor variables showing abrupt process change after 2300 hours in the data sets

Adaptive moving window PLS algorithm

Conventional PLS 알고리즘을 적용했을 때와 동일한 데이터를 사용하여 adaptive moving window PLS 알고리즘을 적용하여 보았다. Initial PLS 모델은 앞의 경우와 동일하게 1000개의 observation set로부터 구하였으며, 모델을 update하기 위한 block size는 100, window size는 1000으로 정하였다. Adaptive 알고리즘을 적용하여 혐기성 공정의 one-step ahead output을 예측한 결과는 Fig 7과 같다. 매번 새로운 100개의 데이터가 들어올 때마다 1000개의 window size를 유지하면서 PLS 모델을 update하였으며, 이때 output quality를 예측하기 위한 optimal latent variable의 수는 cross-validation을 통해 구하였다. 이 경우, PLS model을 update하기 위해서는 사용 가능한 모든 latent variable를 구하는 것이 필수적이므로 optimal LV의 수를 구하는 것과는 별도로, 모델을 update하기 위해서predictor variable의 rank 수만큼 최대한의 latent variable를 구하여야 한다. output을 예측하기 위한 regression coefficient는 이미 구한 latent variable에서 cross-validation을 통해 구한 optimal PC 수까지의 dimension만 이용하면 쉽게 구할 수 있다.

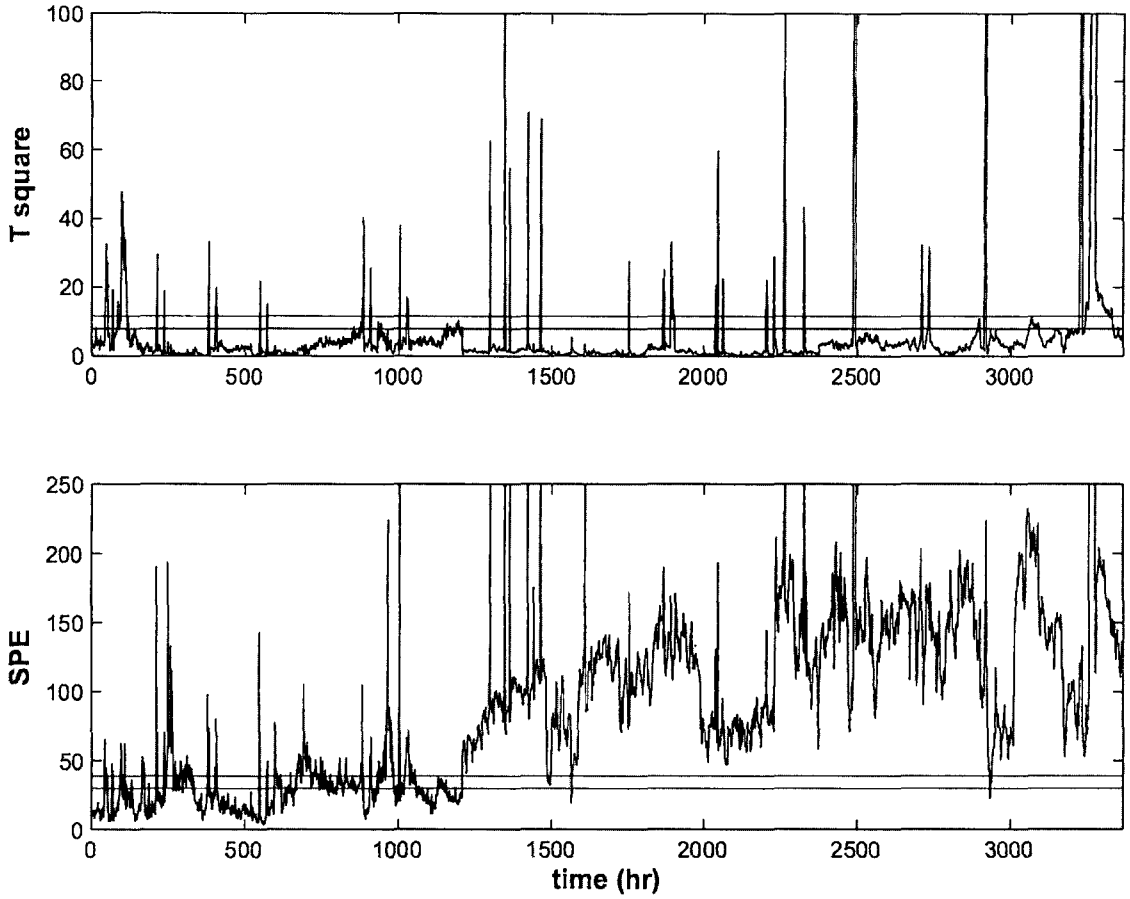

Fig 6. Monitoring chart of conventional PLS model showing process shift in SPE plot within the period of validation data sets

Fig 7의 결과를 살펴보면 conventional PLS 모델에서는 포함되지 않았던 2300시간 대의 correlation structure의 변화가 모델로 update되면서 모델의 예측성능이 보다 향상된 것을 볼 수 있다. 하지만 test data set의 후반부분에 위치한 약 3500시간대 후의 데이터를 시작으로 하여 모델의 예측결과가 상당히 불안정해지면서 모델이 degradation되는 현상을 관찰할 수 있다. 이러한 결과의 원인은 여러 가지로 해석할 수 있다. 우선 첫번째로 생각할 수 있는 문제는 모델이 지속적으로 측정되는 데이터에 존재하는 outlier의 영향을 받아 잘못된 adaptation이 이루어졌을 수 있다는 점이다. 이 경우는 adaptive algorithm 자체의 문제라기보다는 모델을 update할 때 포함되는 새로운 관측 data의 quality와 연관된 문제라고 보아야 하며, 이에 대한 주의가 이루어지지 않을 경우 모델이 fault나 disturbance에 잘못adaptation할 수 있다. 실제 공정에서 측정되는 여러 변수들은 다양한 조건 내에서 sensor fault나 faulty operation, measurement noise 등이 포함되어 있을 수 있기 때문에 모델을 update할 때 사용하기 위한 데이터를 선별할 수 있는 적절한 criteria를 확보하고 filtering의 과정을 거치는 것이 필요하다. 또한 이러한 문제는 adaptive moving window PLS를

실시간으로 적용하였을 때 더욱 심각해 질 수 있으므로 모델의 robustness를 높이기 위한 적절한 방법이 모색되어야 한다. 또 다른 문제는 moving window를 적용함에 있어 이전의 데이터가 계속 새로운 데이터와 바뀌면서 공정의 dynamics를 포함한 중요한 데이터를 잃어버리는 결과를 초래할 수 있다는 점이다.

Fig 7의 모델 예측결과에서 나타나는 blow ups는 약 3400 시간대 부근에서부터 시작되며, 이러한 모델의 destabilization은 모니터링 차트의 confidence limit에도 크게 영향을 미치는 것으로 관찰되었다.

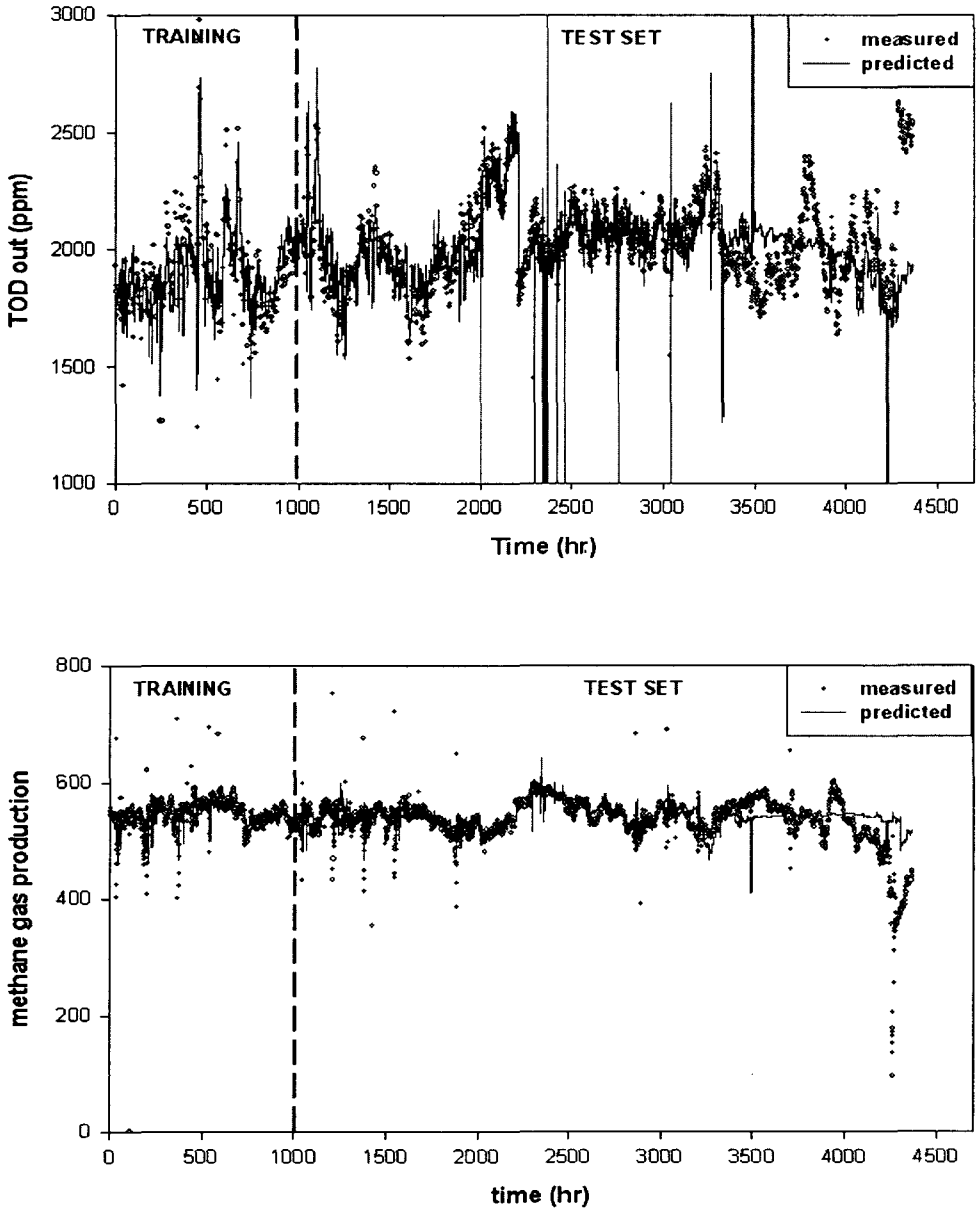

Fig 7. Prediction result using adaptive moving window PLS showing blow ups due to updating the model with faulty observations

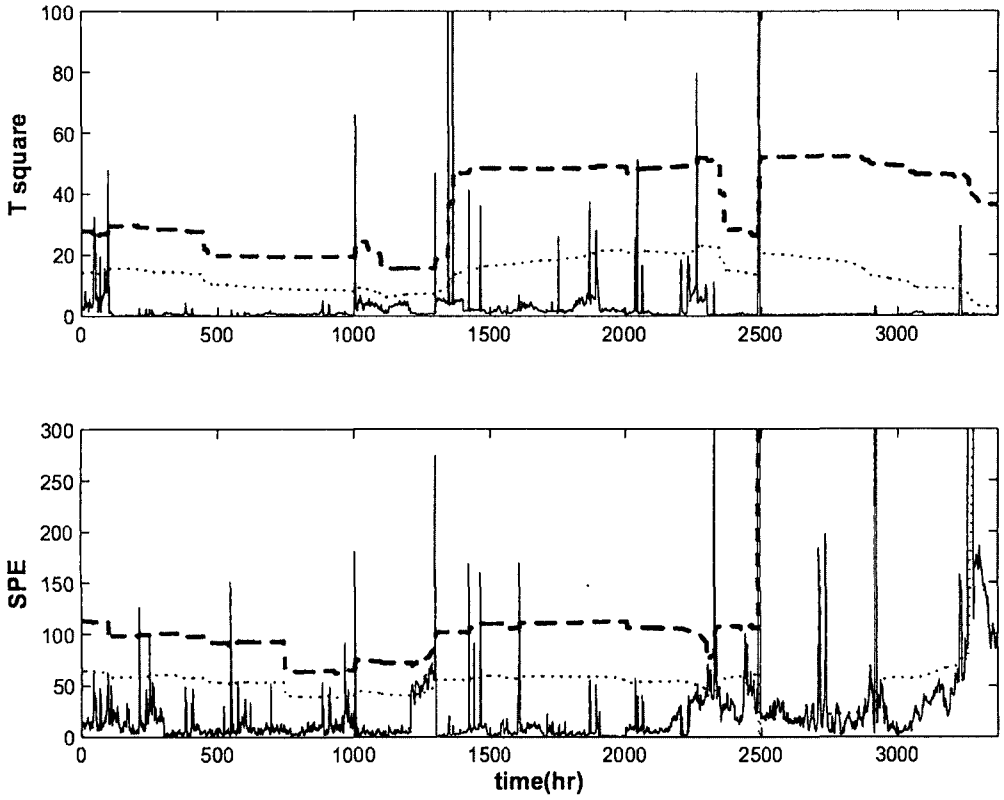

Fig 8. Monitoring chart of adaptive moving window PLS model within the period of validation data sets (short dash line represents 95% confidence limit and long dash represents 99% confidence limit)

특히 이 부근에서 SPE value가 크게 confidence limit를 넘어서며 1000이상의 값을 기록하는 것으로 보아 상당한 정도의 process fault가 일어난 것으로 생각되며, 이 이후부터 SPE값의 fluctuation 정도가 커지는 것을 관찰할 수 있다. 따라서 이러한 변화의 원인을 보다 자세히 조사하여 본 결과, 모델이 불안정하게 된 가장 큰 원인은 pH out의 관측 값에서 생긴 sensor fault임을 확인할 수 있었다. 보통 정상적인 상태에서는 pH out의 측정값이 7 부근에서 거의 일정한 값을 유지하는 데 비해 blow ups가 시작된 시간 근처에서 수 시간동안 지속적으로 pH out의 측정값이 14 부근으로 기록되는 sensor fault가 일어나고 있었으며 그 외에 여러 종류의 minor fault를 포함한 데이터들이 지속적으로 모델로 update됨에 따라 모델이 상대적으로 불안정하게 되는 효과를 가져오게 되었다. 따라서 모델이 abnormal state에 부적절하게 adaptation하거나 outlier에 크게 영향을 받는 문제를 해결하고 모델을 update하기 위한 새로운 데이터를 robust하게 확보하기 위해서는 기존의 update scheme을 개선할 필요가 있음을 위의 결과로부터 확인할 수 있다. 이때 생각할 수 있는 가장 직관적이고 간단한 해결책은 Fig 9와 같은 방법을 사용하는 것이다.

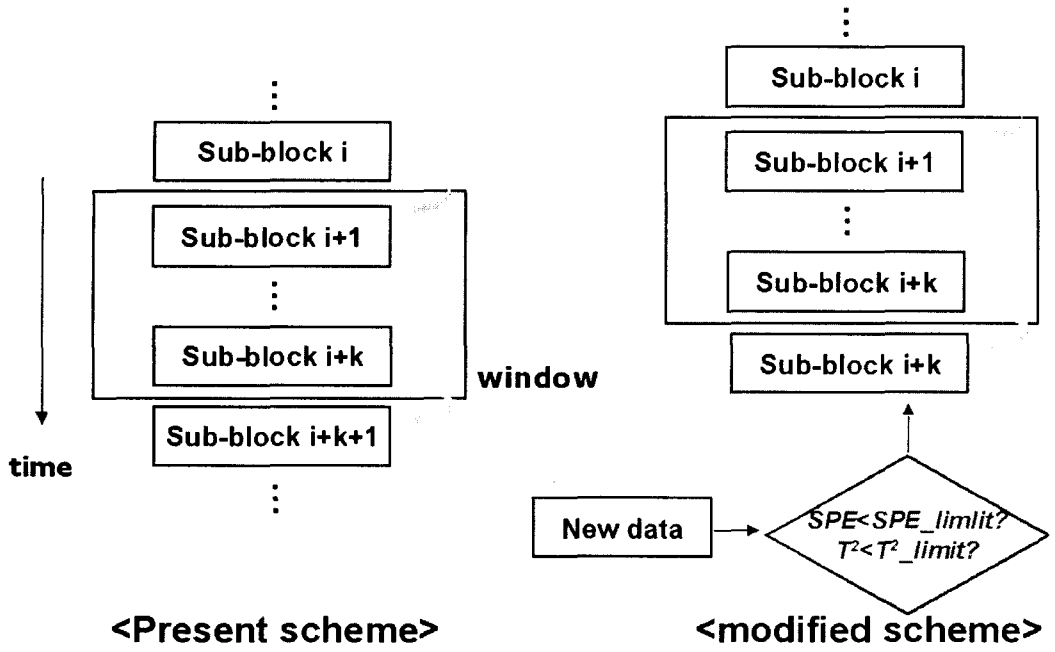

Fig 9. Modified scheme of updating adaptive moving window PLS for improved robustness

새로운 데이터가 지속적으로 들어오면 PLS 모델을 이용한 모니터링을 거친 후, SPE 나 T^2 index가 confidence limit를 넘어설 때를 process fault로 간주하고 모델을 update하기 위한 새로운 block의 데이터에서 포함시키는 과정을 제외시킨다. 위의 방법이 reasonable하게 정당화되기 위해서는 공정에서 일어나는 상태의 변화나 운전조건의 변화가 상대적으로 느리게 일어난다는 가정이 성립하여야 한다. 하지만 이러한 데이터의 rejection scheme은 모델링의 측면에서나 모니터링의 측면에서 볼 때, initial PLS 모델을 만드는 과정에서 매우 routine하게 사용되는 방법이며, 따라서 PLS 모델을 사용하여 output 변수 예측과 공정 모니터링의 목표를 동시에 이루는 데 적용할 수 있는 적당한 방법이다.

앞에서 언급하였던 iteratively reweighted robust PLS의 관점에서 볼 때, 위의 modified scheme은 residual을 monitoring index로 치환한 후, weighting function을 monitoring index가 confidence limit내에 존재할 때 1, limit를 벗어날 때를 0으로 하는 step function으로 정하는 것과 동일한 결과를 가져준다. 이때, SPE 와 T^2 를 동시에 고려하기 위해서 다음과 같은 combined index를 사용하는 것이 가능하다.

$$\varphi = \frac{SPE(x)}{\delta^2} + \frac{T^2(x)}{\chi_l^2} \sim g\chi^2(h) \tag{23}$$

Yue et al.(2001)은 위와 같은 combined index를 사용하여 reconstruction-based fault identification이 가능함을 보여주었으며, 다른 quadratic form의 index들보다 위의 index를 사용함으로써 더 정확하게 variance의 distribution을 설명할 수 있음을 증명하였다. 또한 위의 combined index는 chi-squared distribution을 따르므로 SPE 나 T^2 의 confidence limit를 계산할 때와 같은 방법을 이용하여 쉽게 confidence limit, ξ^2 를 계산할 수 있다.

Robust update of adaptive moving window PLS

모델을 update하는 과정에서 일어나는 blow ups나 outlier의 영향을 방지하고 좀 더 robust하게 모델을 update하는 robust weighted updating scheme은 Table 4와 같이 설명할 수 있다.

Table 4. The adaptive moving window PLS algorithm with robust weighting function

1. Formulate the data matrix $\{X, Y\}$. Autoscale the data
2. Derive a PLS model using the algorithm in Table 1 : $\{X, Y\} \xrightarrow{PLS} \{T, W, P, B, Q\}$
3. When a new pair of data, $\{x_{new}, y_{new}\}$ is available, scale it in the same ways it was done in step 1. Compute combined monitoring index, φ_i , and calculate sample weight as function of φ_i.
4. Multiply sample weight to the data and obtain weighted observation, $\{\omega x_{new}, \omega y_{new}\}$.
5. When a new block of weighted data sets, $\{\omega X_n, \omega Y_n\}$ is available, perform PLS to derive a submodel : $\{X_n, Y_n\} \xrightarrow{PLS} \{T_n, W_n, P_n, B_n, Q_n\}$
6. Formulate $X = \begin{bmatrix} P_i^T \\ P_{i+1}^T \\ \cdots \\ P_n^T \end{bmatrix}$, $Y = \begin{bmatrix} B_i Q_i^T \\ B_{i+1} Q_{i+1}^T \\ \cdots \\ B_n Q_n^T \end{bmatrix}$, and return to step 2. (Discard previous block continuously to maintain the window size)

IRPLS를 adaptive moving window PLS와 함께 사용할 경우, IRPLS의 iteration으로 인해 computational load가 상당히 증가하게 되며, real field에 쉽게 적용하기가 힘든 단점이 있다. 따라서 computation을 줄이고 update 과정을 간단히 하기 위해 본 논문에서는 residual 대신 combined index를 사용하여 weighting function을 구하였으며, one-step iteration of weighting을 통해 combined index의 값이 높은 sample에 대해 weight를 낮게 줌으로써 outlier를 detect하는 방법을 이용하였다. 따라서 이를 통해 모델이 outlier나 abnormal state에 의해 받는 영향을 최소화할 수 있다. 이때 combined monitoring index에 대한 함수로 표현되는 weighting function은 여러 가지의 형태를 취할 수 있으며 이를 정리하면 Table 5와 같다.

Table 5. Functions used for combined monitoring index based weighting algorithms

categories	Weighting function
Hard rejection of the potential outliers	$\omega_i = 1 \ for \ \varphi_i < \xi^2, \ 0 \ for \ otherwise$
Soft rejection of the potential outliers	$\omega_i = \frac{1}{1+(\varphi_i / c\xi^2)^2}$ (Cauchy)
	$\omega_i = \frac{1}{(1+\varphi_i / c\xi^2)^2}$ (Fair)
	$\omega_i = 1 \ for \ \varphi_i \le \xi^2, \ \xi^2 / \varphi_i \ for \ \varphi_i > \xi^2$ (Huber)
	$\omega_i = [(1-\varphi_i^2 / \xi^2)/c^2]^2 \ for \ \varphi_i \le c\xi^2, \ 0 \ for \ \varphi_i > c\xi^2$ (Bisquare)

우선 method with hard rejection of the potential outliers는 지속적으로 update되는 PLS 모델로부터 새로운 샘플의 combined index를 계산하고 이 값이 confidence limit내에 있을 경우에는, 모델을 update하기 위한 block 내로 저장한다. 하지만, combined index가 confidence limit를 벗어날 경우에는 weight를 0으로 줌으로써 이 때를 potential outlier로 간주하고 update 과정에서 제외한다. 이때, weight는 1 또는 0의 값만 가질 수 있으며, 공정의 변화가 느릴 경우에 매우 효과적으로 사용할 수 있다. 하지만 PLS 모델이 공정의 변화를 모델에 update하는 데는 어느 정도의 time-delay가 존재하기 때문에 빠른 시간동안 공정이 변화하거나 새로운 correlation source가 발생하게 될 경우, fault가 일어나지 않았음에도 불구하고 monitoring index가 confidence limit를 넘어서는 결과를 초래할 수 있기 때문에 이를 사용하는 데 있어 주의가 필요하다. Hard rejection 방법을 사용하여 adaptive moving window PLS 모델을 update하고 output을 예측한 결과는 Fig 10과 같다. Fig 10의 결과를 보면 output quality의 예측결과가 outlier를 제거하지 않을 때와 비교하여 상당히 향상되었음을 확인할 수 있으며, blow-ups 현상 또한 나타나지 않음을 확인할 수 있다. 따라서 모델의 예측성능 면에서는 만족할 만한 결과를 얻을 수 있었다고 볼 수 있다. 하지만 adaptive PLS 모델을 변수의 예측 뿐만 아니라 공정의 모니터링에도 동시에 적용하기 위해서는 우리가 얻은 모델이 quality를 적절하게 예측하는가 뿐만 아니라 공정의 변화정도 또한 모델에 얼마만큼 수용할 수 있는가를 동시에 살펴보아야 한다.

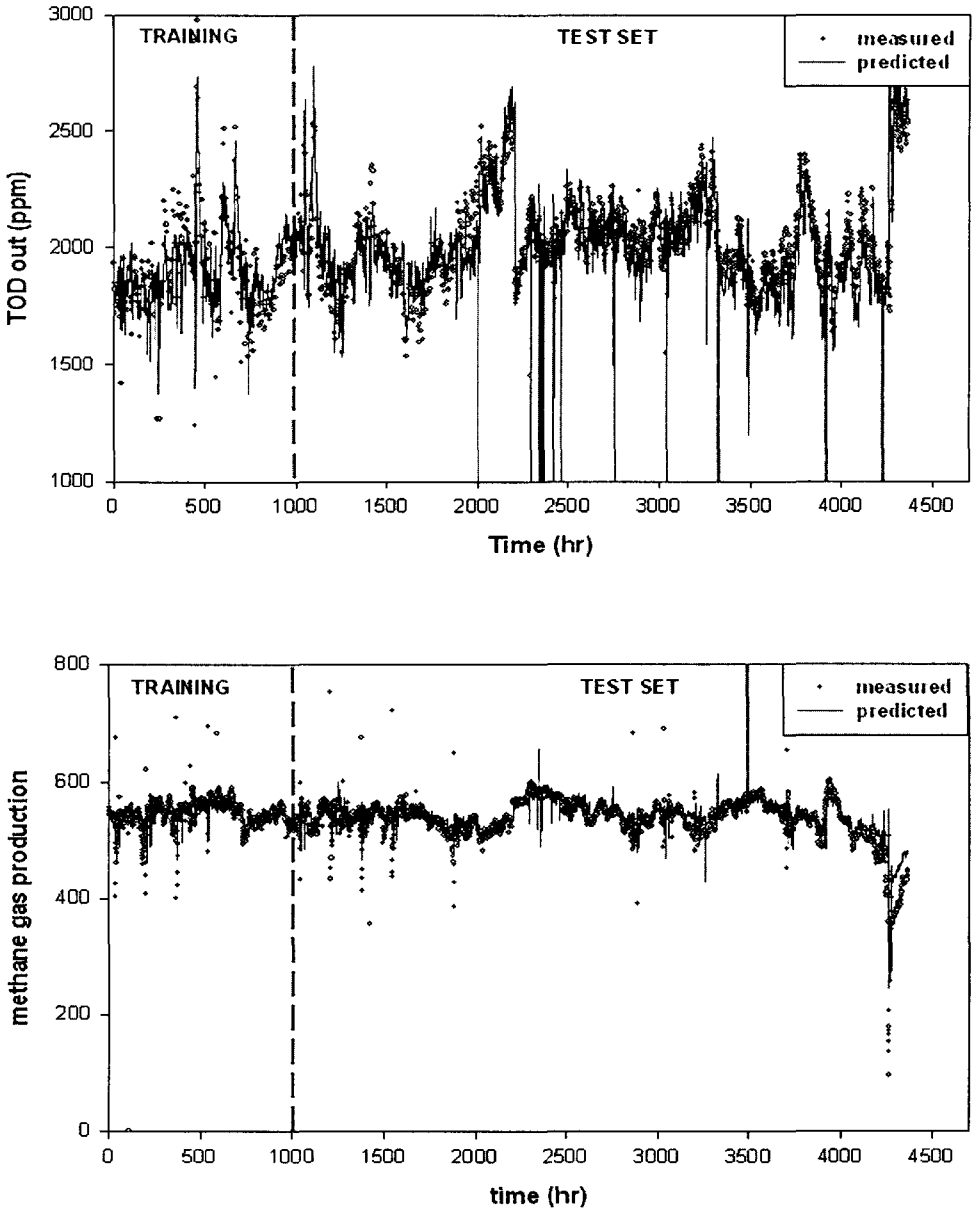

Fig 10. Prediction result using adaptive moving window PLS with hard rejection of the outliers

Model identification을 통해 output quality를 예측하는 것은 regression problem에 속하기 때문에 bias term을 사용하면 공정의 mean, variance의 값이 변화하더라도, correlation structure가 크게 변화하지 않고 충분한 size의 데이터만 확보되면 공정의 변화가 충분히 모델에 update되지 않았더라도 PLS 모델을 통해 얻어진 model의 parameter를 통해 충분한 예측성능을 발휘할 수 있다. 특히 본 논문에서와 같이 충분히 많은 양의 데이터가 존재할 경우 parameter의 variance가 작아지기 때문에 좀 더 정확한 estimation이 가능하게 된다. 하지만 모니터링의 경우, 이러한 공정의 변화를 모델에 포함하지 않게 되면 데이터에 존재하는 variance를 적절히 capture하기가 불가능하기 때문에 monitoring index를 해석하는데 있어 잘못된 결과를 불러올 수 있다.

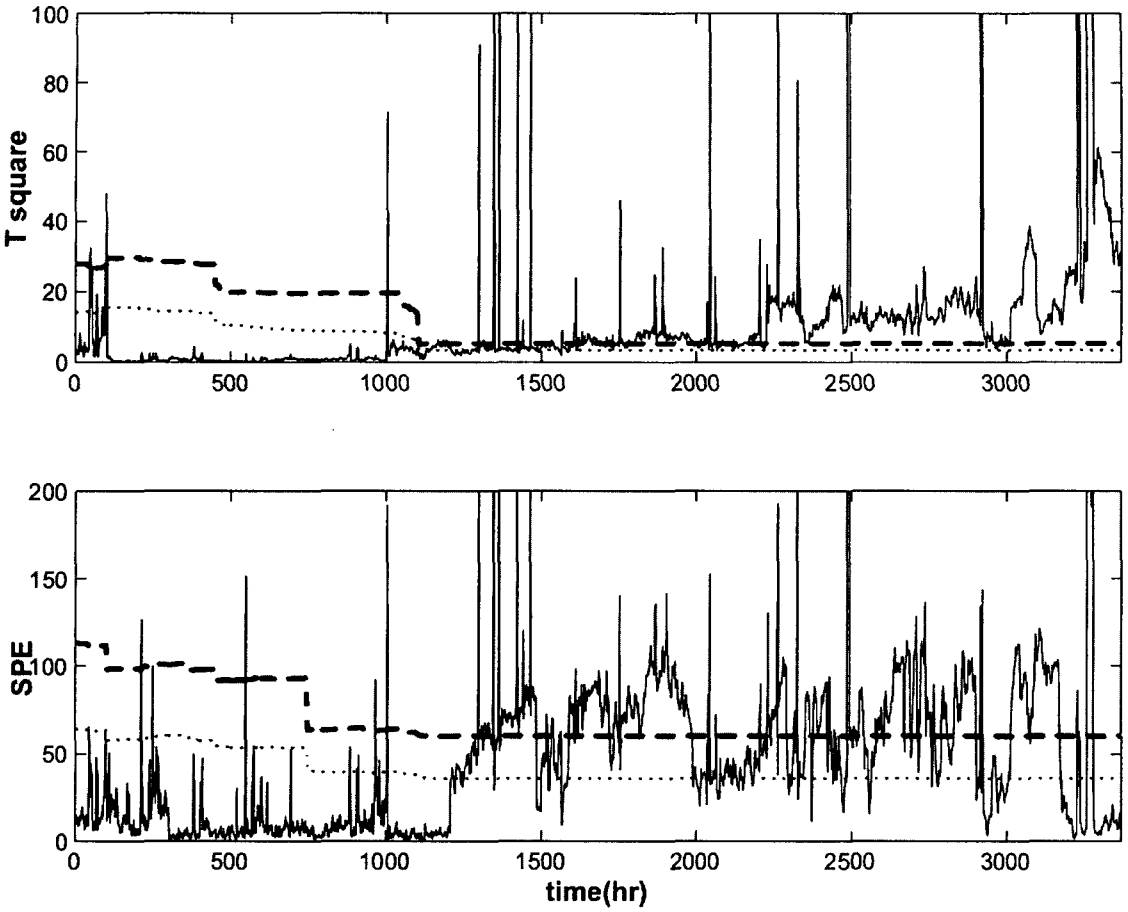

Fig 11. Monitoring chart of adaptive moving window PLS with hard rejection of the outliers within the period of validation data sets (short dash line represents 95% confidence limit and long dash represents 99% confidence limit)

Fig 11은 이러한 결과를 잘 보여주는 것으로 Fig 10을 통해 adaptive model이 output quality를 잘 예측하고 있음을 확인하였음에도 불구하고, 모니터링 차트만을 놓고 볼 때는 공정의 변화가 PLS 모델에 잘 adaptation되지 않았음을 확인할 수 있다. 특히 이러한 변화는 *SPE* chart에서 두드러지게 나타나며 T^2 chart에서는 그 변화가 그다지 크지 않은 것으로 보아 공정의 변화로 인한 영향이 PLS의 모델 space 밖에서 크게 일어나고 있는 것으로 볼 수 있다. 하지만 새로운 correlation source가 발생하였음에도 불구하고 output quality의 예측결과는 크게 영향을 받지 않는 것으로 보아 이러한 correlation의 변화는 X와 Y사이의 correlation structure의 변화가 아닌 오직 X space 상에서만 일어난 공정의 변화로 추정이 된다.

따라서 우리는 hard rejection method가 위와 같은 급작스런 공정의 변화에 적절히 대응하지 못함을 확인할 수 있었다. 이는 공정의 변화나 non-stationary behavior가 모니터링 차트 상에서 confidence limit를 넘게 하는 효과가 있기 때문에 다른 process fault나 abnormal

disturbance를 masking하여 이들간의 구분을 힘들게 하기 때문이었다. 이때, 공정의 변화가 각 변수의 distribution에 주는 영향은 크게 mean, variance vector의 변화와 correlation matrix의 변화, 세 부분으로 나눌 수 있다. 이들 각각의 변화에 따른 monitoring index의 violation은 수식적으로 다음과 같이 설명할 수 있다.

먼저 sample vector, x 를 PCA 모델에 의해 설명되는 부분과 PCA 모델에 의해 설명되지 않는 두 부분으로 나누어진다고 생각하면, 다음과 같이 표현할 수 있다.

$$x = \hat{x} + \tilde{x},\ \hat{x} = PP^T x \in PCS$$
$$\tilde{x} = (I - PP^T)x \in RS \tag{24}$$

또한 sample vector, x 의 SPE 와 T^2 값은 Yue et al(2001)을 참고하여 다음과 같이 구할 수 있다.

$$SPE = \left\|(I - PP^T)x\right\|^2 = \left\|\tilde{x}\right\|^2 \tag{25}$$

$$T^2 = \left\|\Lambda^{-1/2}P^T PP^T x\right\|^2 = \left\|\Lambda^{-1/2}P^T \hat{x}\right\|^2 \tag{26}$$

우선 첫번째로 normal state에서 sample vector, x 의 평균값이 변화하였을 경우, sample vector x 는 다음과 같이 표현되어진다.

$$x = x_{old} + m \tag{27}$$

x_{old} 는 평균값이 변하기 이전의 normal distribution으로부터 구해지는 값이며, m은 평균값의 변화량을 의미한다. 이때 SPE 와 T^2 의 계산은 다음과 같다.

$$SPE = \left\|(I - PP^T)x\right\|^2 = \left\|(I - PP^T)(x_{old} + m)\right\|^2 = \left\|\tilde{x}_{old} + \tilde{m}\right\|^2 \geq \left\|\tilde{m}\right\|^2 - \left\|\tilde{x}\right\|^2 \tag{28}$$

$$T^2 = \left\|\Lambda^{-1/2}P^T PP^T x\right\|^2 = \left\|\Lambda^{-1/2}P^T (PP^T (x_{old} + m))\right\|^2$$
$$= \left\|\Lambda^{-1/2}\mathrm{P}^{\mathrm{T}}(\hat{\mathrm{x}}_{\mathrm{old}} + \hat{\mathrm{m}})\right\|^2 \geq \left\|\Lambda^{-1/2}\mathrm{P}^{\mathrm{T}}\hat{\mathrm{m}}\right\|^2 - \left\|\Lambda^{-1/2}\mathrm{P}^{\mathrm{T}}\hat{\mathrm{x}}_{\mathrm{old}}\right\|^2 \tag{29}$$

fault가 없는 상황에서 $\|\tilde{x}_{old}\|^2$ 와 $\|\Lambda^{-1/2}P^T x_{old}\|^2$ 는 각각 평균값이 변하기 이전의 normal state에서의 SPE 와 T^2 값이며, 따라서 confidence limit에 비해 매우 작은 값을 가진다. 따라서 mean vector의 변화가 monitoring index에 미치는 영향은 m의 값에 따라 좌우된다. 이때 m vector가 PCA loading의 방향과 일치할수록 T^2 의 값이 커지며, 그렇지 않을 경우에는 SPE 의 값이 커지게 된다.

두 번째로 변수의 variance가 변할 경우에도 마찬가지로 이때, x 는 다음과 같이 표현되어진다.

$$x = Ax_{old} \tag{30}$$

이때 A는 각 변수의 variance가 변화하는 ratio를 diagonal element로 하는 square matrix이며, x_{old} 는 첫 번째와 동일하다. 따라서 SPE 와 T^2는 다음과 같이 나타낼 수 있다.

$$SPE = \left\|(I - PP^T)Ax_{old}\right\|^2 = \left\|\tilde{A}x_{old}\right\|^2 \tag{31}$$

$$T^2 = \left\|\Lambda^{-1/2}P^T PP^T Ax_{old}\right\|^2 = \left\|\Lambda^{-1/2}P^T \hat{A}x_{old}\right\|^2 \tag{32}$$

위의 식을 통해 우리는 variance의 변화가 PCA모델에 의해 잘 설명되어지지 않았던 변수에서 크게 일어날 경우에는 SPE 가, 반대로 PCA모델에 의해 잘 설명되었던 변수에 대해 크게 일어날 경우에는 T^2 값이 커지리라는 것을 예상할 수 있다.

마지막으로 변수간의 correlation의 변화가 생겼을 경우를 생각해보면 sample vector, x는 다음과 같이 표현되어진다.

$$x = x_{old} + Cd \tag{33}$$

$C \in \Re^{n \times k}$ 는 이전의 PCA 모델에 의해 capture되었던 correlation 이외의 독립적인 새로운 correlation source을 나타내는 matrix이며, $d \in \Re^{k \times 1}$ 는 score와 같은 개념으로써 x에서 차지하는 C의 크기를 나타낸다. 이때 C는 x에서 PCA 모델의 loading과 consistent한 부분을 제외한 새로운 correlation loading을 나타내므로 $\hat{C}d = PP^T Cd \approx 0$ 라고 볼 수 있다.

$$SPE = \left\|(I - PP^T)(x_{old} + Cd)\right\|^2 = \left\|\tilde{x}_{old} + \tilde{C}d\right\|^2 = \left\|\tilde{x}_{old} + Cd\right\|^2 \geq \left\|Cd\right\|^2 - \left\|\tilde{x}_{old}\right\|^2 \quad (34)$$

$$T^2 = \left\|\Lambda^{-1/2} P^T PP^T (x_{old} + Cd)\right\|^2 = \left\|\Lambda^{-1/2} P^T (\hat{x}_{old} + \hat{C}d)\right\|^2 \cong \left\|\Lambda^{-1/2} P^T \hat{x}_{old}\right\|^2 \quad (35)$$

변수간의 correlation이 변화하였을 경우, sample vector, x의 SPE와 T^2값은 위의 식과 같으며, 이를 통해 correlation의 변화가 주로 SPE 값의 크기에 영향을 미칠 것이라고 예상할 수 있다.

위의 결과들은 PCA 모델을 가정하고 derive했지만, PLS 모델에도 그대로 적용이 가능하며, 따라서 adaptive PLS 모델에서 공정의 변화와 outlier를 구분하는 것이 어렵게 되는 원인으로 설명할 수 있다.

위의 문제를 해결하기 위해서는 들어오는 매 sample 단위로 모델을 update하여 process의 변화에 최대한 빠르게 모델이 adaptation하게 하여야 한다. 하지만 너무 잦은 모델의 update는 불필요한 PLS loading의 회전이 일어나게 하는 원인이 되어 결과적으로 모델이 destabilized 되어 버릴 수도 있다. 따라서 모델의 adaptation speed와 stability 사이에는 trade-off가 존재하며, 이 외에도 모델의 update 주기를 결정하는 데에는 computation load 등과 같은 실질적인 문제도 함께 고려하여야 한다. 따라서 block-wise 알고리즘에서 combined monitoring chart를 이용할 경우, confidence limit의 violation이 공정의 변화로 인한 것인지 abnormal outlier로 인한 것인지 구별하기 위한 적절한 방법이 필요하며, 이를 위해 wavelet analysis를 통해 change of the process dynamics를 de-trend하고 이를 모니터링하는 연구가 현재 진행중이다.

위에서 살펴본 바와 같이 Hard rejection의 경우, process dynamics로 인한 변화와 abnormal outlier를 구분하는 것이 어려우므로 모델의 update가 제대로 이루어지지 않을 수 있으며, X변수와 Y변수사이에 correlation이 변화할 경우, output quality에 대한 예측성능도 보장할 수 없게 된다. 따라서 새로운 데이터가 들어올 때, 적절한 weight를 줘서 fault와 non-stationary property를 동시에 지닌 실시간 데이터로부터 block size에 해당하는 만큼의 time delay로 인해 모델이 adaptation하지 못했거나 또는 짧은 시간동안 빠르게 일어났던 process dynamics의 변화를 부분적으로 모델에 포함시키면서 severe outlier를detect할 수 있는 method of soft rejection of the outlier의 알고리즘을 사용하는 것이 robust model을 개발하기 위한 보다 더 적절한 방법이라고 할 수 있다.

이때 사용할 수 있는 weighting function의 종류는 Wakeling et al(1992)이나 Commins et al(1995) 등을 참고하여 Table 5와 같이 선정하였다. 이들 각각의 weighting function은 robustness의 정도나 대상 데이터의 범위를 조절할 수 있는 tuning parameter를 가지고 있으며, 이들의 값은 heuristic하게 정하게 된다. 위의 weighting function을 사용하여 output quality를 예측한 결과는 Table 6와 같으며, 이를 실제로 적용할 경우 tuning parameter의값에 따라

예측성능이 상당히 변화하는 결과를 보여주었다. Table 6의 결과는 각 weighting function 별로 tuning parameter를 prediction과 모니터링을 동시에 만족하는 범위 내에서 optimize한 결과이다. 예측성능 면에서 볼 때, 각 방법들 간에 비교적 큰 차이는 없었으나 Fair weighting function을 썼을 때의 방법이 가장 나은 결과를 보이고 있었다.

Table 6. Prediction results using various PLS approaches in the test data sets

		PLS	adaptive PLS					
			no rejection	hard rejection	Cauchy	Fair	Huber	Bisquare
RMSE	TOD out	331.7	244.8 (178.2)*	181.3	185.7	176.3	188.9	183.6
	Q CH4	32.1	38.4 (17.9)*	28.5	26.6	27.9	26.2	27.0
R^2	TOD out	-0.396	0.240 (0.515)*	0.583	0.563	0.605	0.547	0.572
	Q CH4	0.430	0.183 (0.513)*	0.549	0.608	0.566	0.620	0.595

* ( ) values indicate prediction result before the "blow up" has been observed

Table 6의 결과를 좀 더 자세히 분석해보면, weighting function을 사용함으로 인해 blow up가 나타났던 adaptive PLS에 비해 그 예측성능이 훨씬 개선되었음을 확인할 수 있었으며, hard rejection scheme을 썼을 때의 예측성능과 비교했을 때는 그 결과는 거의 비슷하거나 조금 개선되었음을 알 수 있다. 또한 모니터링의 측면에서 볼 때, hard rejection scheme의 PLS 모델에서 나타났던 update의 문제가 모든 weighting function의 방법에서는 나타나지 않음을 확인할 수 있었다. 따라서 위의 결과는 soft rejection scheme의 방법을 통해 output quality와 모니터링의 두 가지 목표를 동시에 이룰 수 있음을 보여주고 있다.

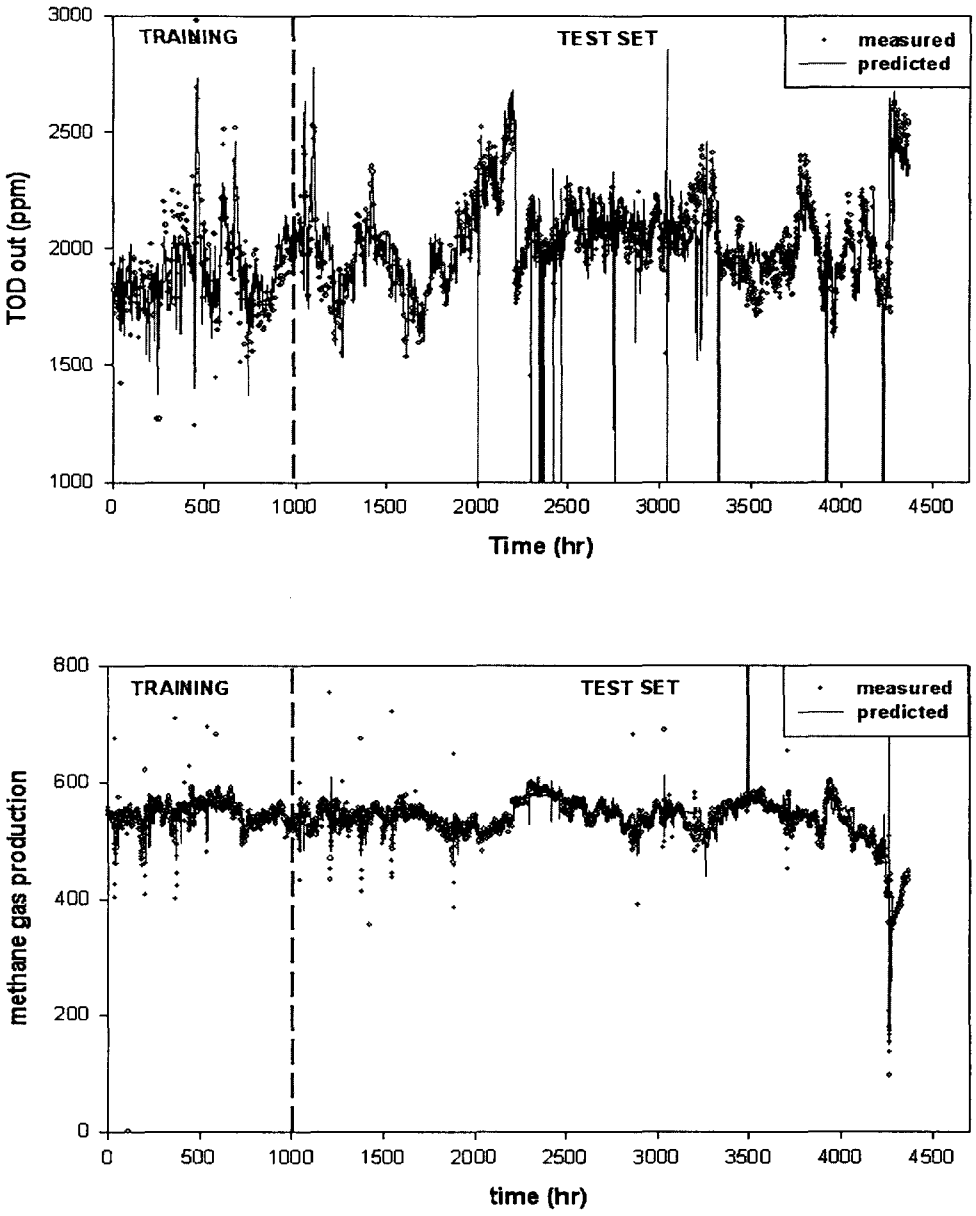

Fig 12. Prediction result using adaptive moving window PLS with Fair weighting function

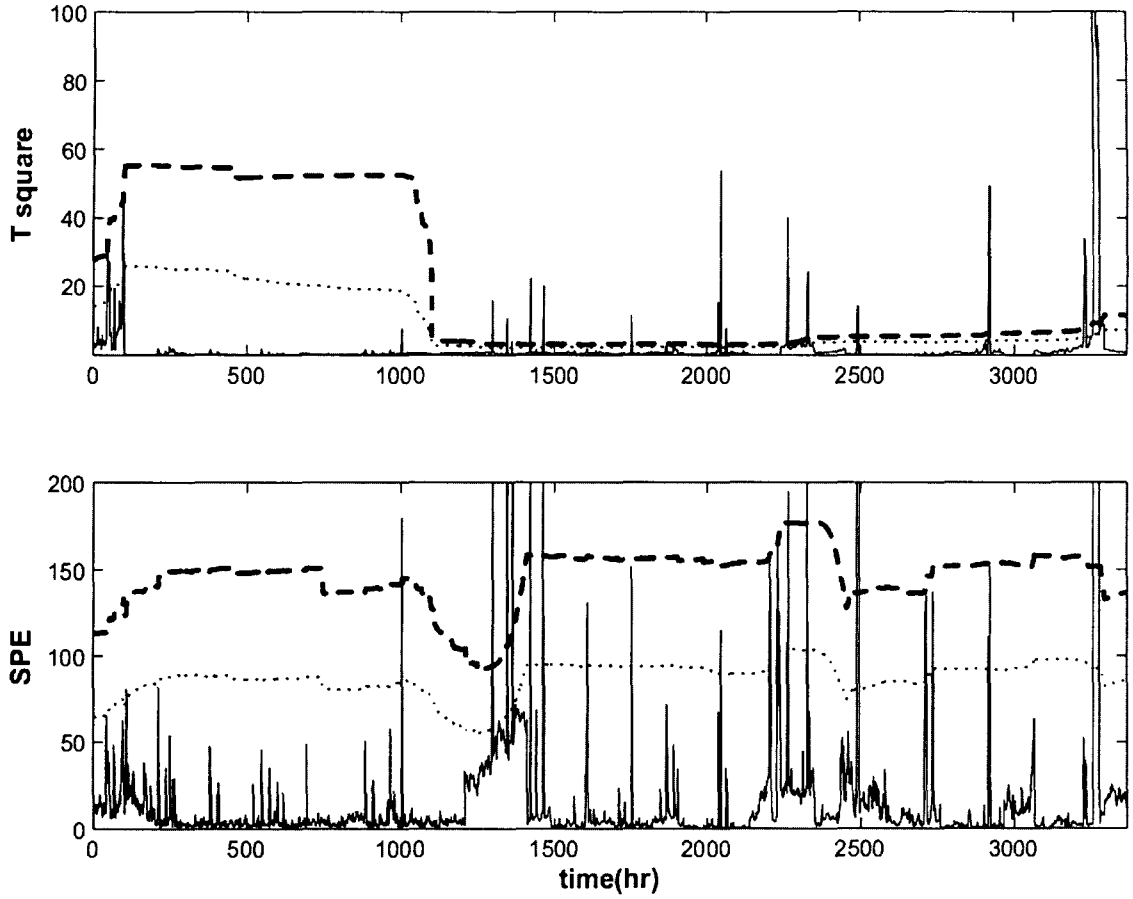

Fig 13. Monitoring chart of adaptive moving window PLS with Fair weighting function within the period of validation data sets (short dash line represents 95% confidence limit and long dash represents 99% confidence limit)

Fig 12와 13은 Fair weighting function을 썼을 때의 prediction과 모니터링 결과를 나타낸 것이다. 특히 Fair weighting function을 써서 outlier를 detection하고 그 영향을 제거하였을 경우, tuning parameter의 값에 따라 예측성능이 Table 6의 결과보다 더욱 향상되는 결과를 보여주었다. 하지만 이 경우, 모니터링 차트에서 Fig 11과 같은 bias가 관찰되었기 때문에 이를 제외하였다. 이러한 원인으로는 현재, 변수의 mean의 변화가 관계되어 있지 않는가 하고 추측하고 있다. 본 논문에서 사용한 adaptive PLS 알고리즘의 경우, 특별히 변수의 pretreatment scaling 단계에서 mean vector를 update하는 대신, training set에서 사용한 값을 그대로 이용하여 scaling을 수행하는 것을 기본으로 하였다. 하지만 이 경우, mean vector가 변화함에 따라 latent space의 shift가 일어나며, 이를 모델에 포함하기 위해서는 additional factor를 사용하는 것이 필요하다.[Helland, 1992] 따라서 모델의 adaptive updating 단계에서 optimal PC의 수를 정할 때, 이를 고려하여 PC의 수를 추가시키는 과정이 포함시키지 않을 경우, mean shift를 설명하는 additional PC가 residual space상에서 점점 증가하여 Fig 11과 같이 SPE 값에 bias가 관찰될 수 있다. 하지만 자동적으로

모델이 update되면서 mean vector가 shift하였는지 확인하고 이를 설명하기 위해 PC를 추가할 것인지를 결정하는 문제는 쉬운 일이 아니다. 따라서 이를 보완하기 위해서는 adaptive algorithm에서 실시간으로 mean이나 bias를 check하고 이에 대응하여 적절하게 PC를 추가하는 방법에 대한 좀 더 자세한 연구가 필요할 것이다.

6. 결 론

본 논문에서는 혐기성 폐수처리 공정의 output quality를 예측하고 공정을 모니터링하기 위한 모델을 만들기 위해 PLS 방법을 이용하였다. 미생물이 포함된 폐수처리 공정은 stationary state에서 운전되는 경우가 거의 없으며, 계절이나 유입 폐수의 성상에 따라 미생물의 활성과 분포가 변화하기 때문에 일정한 covariance matrix로부터 얻어지는 conventional PLS 방법을 적용할 경우, 잘못된 결과를 보일 수 있음을 확인하였다. 또한 시간에 따라 모델이 자동적으로 계속 update되는 adaptive moving window PLS를 혐기성 공정의 실시간 데이터에 적용하고 이를 통해 output quality를 예측할 수 있는 dynamic model을 얻을 수 있었다.

최근 생물학적 폐수처리공정에 다변량 통계분석 기법을 이용하여 효과적으로 공정을 모델링하고 이를 통해 disturbance가 비교적 큰 유입수의 성상으로부터 안정적으로 공정을 감시하고 제어하려는 연구들이 많이 이루어지고 있다. 하지만 industrial anaerobic wastewater treatment plant와 같이, 민감하게 반응하는 미생물의 활성에 따라 operation strategy가 변화하고 정기적인 shut-down이나 시설교체 등이 잦으며, 유입수의 조건을 일정하게 유지하기 힘든 공정의 경우, PLS 모델의 long-term application을 위해서는 adaptive 모델을 이용하는 것이 필수적이다. 따라서 이를 실공정에 적용하기 위해서는 많은 연구와 노력이 필요하다. 본 논문에서는 실공정의 데이터를 이용하여 이러한 issue 들 중에 하나로 생각될 수 있는 robust model update scheme에 대해 연구하여 보았으며, 이를 해결하기 위한 적적한 방법으로 weighting scheme을 제안하였다. Robust한 PCA/PLS 모델을 얻기 위해서 iteratively reweighted algorithm 이외에도, minimum volume ellipsoid (MVE), minimum covariance determinant (MCD), ellipsoidal multivariate trimming (MVT) 등의 여러 가지 방법들이 제안되었지만, adaptive model에 이를 적용하기에는 너무 복잡하거나 high computational loading이 요구되기 때문에 실공정에 적용하기 어려운 점들이 많이 있다. 따라서 본 논문에서 제시한 monitoring index를 이용한 one-step iteration of weighting의 방법을 사용하여 효과적으로 abnormal outlier를 제거함으로써 안정적으로 공정의 output quality를 예측하고 모니터링할 수 있으며, 실제 데이터를 이용하여 직접 이를 확인하여 보았다.

사 사

본 연구는 한국과학재단 지정 포항공대 차세대바이오환경기술연구센터(AEBRC)의 연구비지원으로 수행되었으며 이에 감사드립니다.

참고문헌

Baston DJ, Keller J, Angelidaki I, Kalyuzhnyi SV, Pavlostathis SG, Rozzi A, Sanders WTM, Siegrist H, Vavilin VA. 2002. Anaerobic Digestion Model No. 1(ADM1). IWA Task Group for Mathematical Modeling of Anaerobic Digestion Processes. IWA Publishing, London

Bernard O, Hadj-Sadok Z, Dochain D, Genovesi A, Steyer JP. 2001. Dynamical model development and parameter identification for an aerobic wastewater treatment process. Biotechnol Bioeng 75:424-438

Dayal BS, MacGregor JF. 1996. Identification of finite impulse response model: Method and robustness issues. Ind Eng Chem Res 1996:4078-4090

Dayal BS, MacGregor JF. 1997. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. J Process Control 7:169-179

De Jong S. 1993. SIMPLS: an alternative approach to partial least squares regression. Chemom Intell Lab Syst 18:251-263

Geladi P, Kowalski BR. 1986. Partial least-squares regression: a tutorial. Anal Chim Acta 185:1-17

Helland K, Berntsen HE, Borgen OS, Marten H. 1991. Recursive algorithm for partial least squares regression. Chemom Intell Lab Syst 14:129-137

Jackson JE. 1991. A user's guide to principal components. Wiley-interscience, New York

Juricek BC, Seborg DE, Larimore WE. 2002. Identification of multivariable, linear, dynamic model: Comparing regression and subspace techniques. Ind Eng Chem Res 41:2185-2203

Kourti T, MacGregor JF. 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemom Intell Lab Syst 28:3-21

Ku W, Storer RH, Georgakis C. 1995. Disturbance detection and isolation by dynamic principal component analysis. Chemom Intell Lab Syst 30:179-196

Li B, Morris J, Martin EB. 2002. Model selection for partial least squares regression. Chemom Intell Lab Syst 64:79-89

Li W, Yue HH, Valle-Cervantes S, Qin SJ. 2000. Recursive PCA for adaptive process monitoring. J Process Control 10:471-486

Lee DS, Jeon CO, Park JM, Chang KS. 2002. Hybrid neural network modeling of a full-scale industrial wastewater treatment process. Biotechnol Bioeng 78:670-682

Lee DS, Vanrolleghem PA. 2002. Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis. Biotechnol Bioeng 82:489-497

Lennox J, Rosen C. 2002. Adaptive multiscale principal component analysis for online monitoring of wastewater treatment. Wat Sci Technol 45:227-235

Lübbert A, Jørgensen SB. 2001 Bioreactor performance: a more scientific approach for practice. J Biotechnol 85:187-212

MacGregor JF, Kourti T, 1995. Statistical process control of multivariate processes. Control Eng Practice 3:403-414

Marsili-Libelli S, Müller A. 1996. Adaptive fuzzy pattern recognition in the anaerobic digestion process. Pattern Recogn Lett 17:651-659

Mujunen SP, Minkkinen P, Teppola P. Wirkkala RS. 1998. Modeling of activated sludge plants treatment efficiency with PLSR: a process analytical case study. Chemom Intell Lab Syst 41:83-94

Osten DW. 1988. Selection of optimal regression model via cross-validation. J Chemom 2:39-48

Qin SJ. 1998. Recursive PLS algorithms for adaptive data modeling. Comput Chem Eng 22:503-514

Rosen C, Olsson G. 1998. Disturbance detection in wastewater treatment plants. Wat Sci Technol 37:197-205

Rosen C, Lennox JA. 2001. Multivariate and multiscale monitoring of wastewater treatment operation. Wat Res 35:3402-3410

Shi R, MacGregor JF. 2000. Modeling of dynamic systems using latent variable and subspace methods. J Chemom 14:423-439

Steyer JP, Buffière P, Rilland D, Moletta R. 1999. Advanced control of anaerobic digestion processes through disturbances monitoring. Wat Res 33:2059-2068

Steyer JP, Genovesi A, Harmand J. 2001. Advanced monitoring and control of anaerobic wastewater treatment plants: fault detection and isolation. Wat Sci Technol 43:183-190

Tay JH, Zhang X. 2000. A fast predicting neural fuzzy model for high-rate anaerobic wastewater treatment systems. Wat Res 34:2849-2860

Teppola P, Mujunen SP, Minkkinen P. 1997. Partial least squares modeling of an activated sludge plant: A case study. Chemom Intell Lab Syst 38:197-208

Teppola P, Minkkinen P. 2000. Wavelet-PLS regression models for both exploratory data analysis and process monitoring. J chemom 14:383-399

Tomita RK, Park SW, Sotomayor OAZ. 2002. Analysis of activated sludge process using multivariate statistical tools- a PCA approach. Chemical Engineering Journal 90:283-290

Vijaysai P, Gudi RD, Lakshminarayanan S. 2003. Identification on demand using blockwise recursive partial least-squares technique. Ind Eng Chem Res 42:540-554

Wakeling IN, Macfie JHH. 1992. A robust PLS procedure. J chemom 6:189-198

Wang X, Kruger U, Lennox B. 2003. Recursive partial least squares algorithms for monitoring complex industrial processes. Control Eng Practice 11:613-632

Wise BM, Ricker NL. 1990. The effect of biased regression on the identification of FIR and ARX models. AIChE annual meeting???

Wise BM, Gallagher NB. 1996. The process chemometrics approach to process monitoring and fault detection. J Process Control 6:329-348

Wold S. 1978. Cross-validatory estimation of the number of components in factor and principal component analysis. Technometrics 20:397-405

Wold S. 1994. Exponentially weighted moving principal component analysis and projections to latent structures. Chemom Intell Lab Syst 23:149-161

Wu TJ, Sepulveda A. 1998. The weighted average information criterion for order selection in time series and regression models. Stat Probab Lett 39:1-10

Yue HH, Qin SJ. 2001. Reconstruction-based fault identification using a combined index. Ind Eng Chem Res 40:4403-4414