HWAHAK KONGHAK Vol. 41, No. 5, October, 2003, pp. 592-597

블록들의 유사성을 고려한 Adaptive Block-Wise RPLS

윤경우 · 이영학 · 한종훈*,†

포항공과대학교 환경공학부, *화학공학과 및 아이시스텍(주)
790-784 경북 포항시 남구 효자동 산 31
(2003년 1월 13일 접수, 2003년 6월 25일 채택)

Adaptive Block-Wise RPLS Considering Similarity of Blocks

Kyong-U Yun, Young-Hak Lee and Chonghun Han*,†

School of Environmental Science and Engineering,
**Department of Chemical Engineering, Pohang University of Science and Technology and ISYSTECH Co. Ltd.*
San 31, Hyoja-dong, Nam-gu, Pohang, Kyungbuk 790-784, Korea
(Received 13 January 2003; accepted 25 June 2003)

요 약

부분 최소 자승법(partial least squares, PLS)은 복잡한 공분산 구조를 가진 상관성이 큰 변수들이 포함된 공정을 모델링 하는데 효과적으로 사용되어 왔다. 하지만 공정 변화에 적응하는 온라인 모델을 만들기 위해 부분 최소 자승법을 이용하는 것은 많은 제약이 따른다. 그래서 많이 이용되고 있는 방법이 recursive PLS(RPLS)이다. RPLS는 새로운 데이터 샘플이 들어오면 모델을 갱신 한다. 하지만 데이터의 상당한 양이 모이거나 공정의 큰 변화가 있을 때까지 모델을 갱신하지 않을 수도 있다. 따라서 이 논문에서는 공정 변화에 잘 적응하고 모델 갱신을 블록 단위로 하는 block-wise RPLS의 온라인 적응 방법인 moving window 방법을 이용하였다. Block-wise RPLS는 회귀 모델을 만들고 검증하는데 필요한 계산 부하를 감소시키는 효과가 있다. 본 연구에서는 모델을 구성하는 윈도우내에 있는 블록들 사이의 상관관계를 나타내는 지표를 정의한 다음, 그 지표를 통해 forgetting factor를 정하는 방법을 제시하였다. 이 방법은 공정에 대한 일반적인 지식이 필요하지 않고 단지 데이터 블록의 상관관계만으로 적응 모델의 예측력을 높일 수 있는 방법이다. 따라서 다른 방법들이 나타내지 못한 상관성 정보를 포함시켰다. 이 방법을 기반으로 정유 공장의 가열로의 NO_x(nitrogen oxides) 배출 농도를 예측하는 모델을 forgetting factor를 이용해 구성했고, 이 forgetting factor에 의해 예측력이 최소 에러 값의 95%이상 근접함을 확인하였다.

Abstract – Partial least squares (PLS) regression has been effectively used as one of the data-driven empirical modeling to deal with a large number of variables in the complicated covariance structure. But Partial least squares (PLS) regression has limitations in the online model update. Therefore, recursive PLS has been used for the online adaptation of the model. The RPLS algorithm is implemented as soon as some new samples are available. It may be desirable not to update the model until significant amount of data are collected and the process has gone through significant changes. In this paper, we used the block-wise recursive partial least squares (RPLS) algorithms with a moving window for the adaptation of the model according to the process shifts. The block-wise RPLS algorithm has been used to reduce computational load in PLS regression and its cross-validation. In this work we defined the index to represent correlation between blocks and proposed how to determine the forgetting factors through it. This proposed method is to improve the prediction power of the adaptive modeling considering correlation between blocks without general process knowledge. Therefore, this proposed method included correlation information that different methods did not express. The method was tested to process heaters in the oil refining company. We constructed the prediction model of the effluent NO_x composition with forgetting factors and showed that the prediction power approximated more than 95% of the minimum error.

Key words: Partial Least Squares, Recursive PLS, Block-Wise RPLS, Forgetting Factors, Moving Window, Adaptive Modeling, Correlation, Heater, NO_x

†To whom correspondence should be addressed.
E-mail: chan@postech.ac.kr

1. 서 론

공정의 과거 조업 데이터만 있으면 적용 가능한 다변량 통계 분석 방법 중 하나인 PLS(partial least squares)는 선형 회귀 방법들 중 가장 뛰어난 성능을 나타낸다. Wold가 선구적인 업적을 남긴 이래로, PLS는 chemometrics 분야에서 널리 적용되고 있다[1-3]. PLS 회귀법은 더 고전적인 회귀법인 MLR(multiple linear regression)이나 PCR(principal component regression)보다 더 강건한 모델을 제공하기 때문에 좋은 대안으로서 사용된다. PLS 응용의 대부분에서, PLS 회귀법은 batch-wise 모델링적 접근을 사용한다. 즉, 컴퓨터에 데이터가 모아지고 저장된 다음, PLS 회귀법이 전체 데이터 영역에 의해 수행된다. Batch-wise PLS 회귀법은 공선형성(collinearity) 문제를 피할 수 있지만, 다음과 같은 세 가지의 제약을 가지고 있다. 첫째, 새로운 데이터가 들어 왔을 때, 온라인상에서 PLS 모델을 갱신하기가 어렵다. 즉, 새로운 데이터가 들어오면 새로운 데이터와 과거 데이터가 합쳐져서 새로운 모델을 형성하게 된다. 이것은 과거 데이터를 재사용해서 모델이 형성되므로, 계산상으로 비효율적이다. 둘째, 많은 변수와 샘플을 가진 데이터일 경우, batch-wise PLS 알고리듬은 컴퓨터 메모리를 다 써버릴 수도 있다. 셋째, cross-validation을 수행할 때, 너무 많은 시간이 소요된다[4]. 이러한 문제들을 해결하기 위해 사용된 방법이 RPLS(recursive PLS)이다. 기본적인 RPLS 알고리듬은 Helland 등[5]에 의해 제안되었으며, Qin[4]에 의해 수정되고 moving window와 forgetting factor를 이용한 block-wise RPLS 방법으로 확장되었다. 그리고 Qin[4]은 새로운 데이터에 대한 서브 모델과 과거 PLS 모델을 결합해서 새로운 모델을 만드는 block-wise RPLS 방법으로 확장했고 cross-validation 절차에도 적용하였다. 이 방법은 기존의 batch-wise PLS보다 계산 부하가 적어서 시간과 메모리 측면에서 장점을 갖고 있고 온라인상에서 모델 갱신이 용이하다. 그리고 PLS 알고리듬의 계산 속도를 빠르게 하기 위해 개선된 kernel 알고리듬을 이용한 RPLS 방법도 제안되었다[6]. 본 연구에서 사용된 기본적인 방법은 Qin[4]에 의해 제안된 block-wise RPLS with a moving window and forgetting factors 방법이다. 이 방법에서 모델의 예측력을 높이기 위해 사용되는 중요한 변수는 forgetting factors이다. 본 논문에서는 forgetting factor를 정하는 방법을 제안하였다.

본 저자가 제안한 방법을 실제 정유공장의 가열로에서 배출되는 NO_x 농도 모델링에 적용하였다. 최근 들어 환경 문제가 전 세계적인 문제로 대두되어 환경 공정에 대한 데이터 분석, 모니터링, 모델링과 제어에 관한 연구가 많이 발표되었다[7-10]. 이러한 연구들은 퍼지, 신경망, 유전자 알고리듬, 회귀 모델 등 다양한 방법을 통해 연구되었다. 본 저자가 적용한 가열로는 온라인 분석기에 의해 NO_x 농도를 표시하는데, 온라인 분석기의 오작동에 대비하고 보다 빠른 NO_x 배출 농도의 제어를 위해 정확한 NO_x 배출 농도 모델이 필요했다. 따라서 공정의 변화에 적응하는 정확한 모델을 유지하기 위해 온라인 모델 갱신이 필수적이었고 모델의 예측력을 높일 수 있는 방법으로 forgetting factor의 결정에 초점을 두었다. Forgetting factor는 모델에 사용된 블록의 상관관계를 나타내는 저자가 제안한 지표에 의해 쉽게 결정되었고, 모델링에 사용되는 자료 행렬을 구성하는데 이용되었다.

2. 이론 및 제안한 방법론

2-1. Partial Least Squares Regression

X와 Y, 두 자료 행렬 사이의 선형 관계를 분석하는데 있어서 주로 사용되어 온 방법은 다중 선형 회귀법(multiple linear regression, MLR)이다. 그러나 각 행렬 내에 서로 상관성이 큰 변수들이 포함된 경우, 그 예측력은 상당히 떨어지게 되는데 PLS는 이 문제를 효과적으로 처리하여 예측력이 뛰어나고 공정 잡음과 어느 정도의 센서 고장 시에도 강건한 모델을 제공할 수 있는 방법이다. PLS 모델을 만들기 위해 우선 모니터 되어야 할 응답 변수들, 즉 품질 변수들의 데이터들로 Y를 구성하고, 예측자 변수들, 즉 온라인으로 측정되는 공정 변수들로 X를 구성한다. 입력(X)과 출력(Y)에 대한 자료 행렬 한 쌍이 주어지면, 두 행렬을 다음과 같은 선형 관계로 가정한다.

$$Y=XC+V \tag{1}$$

여기서 V와 C는 각각 noise 행렬과 회귀 계수(regression coefficient) 행렬을 나타낸다. 두 자료 행렬, X와 Y의 상관관계를 이용하여 자료 행렬을 분해하고, 선형 모델을 만든다. 그 결과로서 형성되는 행렬은 다음과 같다.

$$X=TP^T+E \tag{2}$$
$$Y=UQ^T+F \tag{3}$$
$$U=TB+G \tag{4}$$
$$B=(T^TT)^{-1}T^TU \tag{5}$$

여기서 E, F, G는 잔차(residual) 행렬들이고 T는 Y를 고려한 X 축소 공간에서의 이력을 나타내는 score 행렬, U는 X를 고려한 Y 축소 공간에서의 이력을 나타내는 score 행렬이다. 그리고 P와 Q는 각각 X와 Y에 대한 loading 행렬이고, B는 X와 Y사이의 내적 관계를 표현하는 내적 모델 계수들의 diagonal 행렬이다. 위의 행렬들을 구하기 위해 Geladi와 Kowalski[11]에 의해 소개된 PLS 알고리듬을 Table 1에 나타내었다. 모델에서 factors의 수를 결정하는 방법은 F-test가 제안되기도 하였지만[11], cross-validation(CV)이 가장 기본적이고 실제적이면서도 신뢰할 수 있는 방법이다[12]. 기본적으로 CV는 데이터를 많은 그룹들(예를 들면 5-10개의 그룹들)로 나눈 후에 그룹들 중에서 한 그룹을 제외시킨 데이터를 여러 세트로 만들게 된다. 이 세트들로 각각 PLS 모델을 만든 후에 각 모델을 만들 때 제외된 그룹이 각 모델의 검증 그룹이 되어 이 검증 그룹에 대한 Y 변수들의 실제 값과 예측 값의 차이를 계산한다. 모든 모델로부터 이 차이들의 SS(sum of squares)를 계산하고 합산하게 되면 그 모델의 예측력에 대한 척도인 PRESS (predictive residual sum of squares)가 된다. 이 값이 거의 최소가 되고 이 값과 latent variables(LV)의 그래프에서 기울기가 거의 0에 가까운 지점에서 LV의 수, 즉 factor의 수를 정하게 된다. 보통 PRESS는 Q^2(the "cross-validated R^2")로 재표현 되는데 이 값은 $(1-PRESS/SS_y)$이며 여기서 SS_y는 Y의 평균에 대한 Y의 SS이다. 이 값은 $R^2=(1-RSS/SS_y)$와 비교할 수 있는데 이 R^2는 0에서 1사이의 값을 가지며 1은 완전한 모델임을, 0은 전혀 관련이 없는 모델임을 나타낸다. 여기서 RSS(residual sum of squares)는 모델이 설명할 수 없는 residual들의 SS이다. 일반적으로 Q^2도 0과

Table 1. A traditional batch-wise PLS algorithm

1. Scale **X** and **Y** to zero-mean and unit-variance.
   Initialize $\mathbf{E}_0:=\mathbf{X}$, $\mathbf{F}_0:=\mathbf{Y}$, and h:=0.
2. Let h:=h+1 and take $\mathbf{u}_h$ as some column of $\mathbf{F}_{h-1}$.
3. Iterate the PLS outer model until it converges:
   $\mathbf{w}_h=\mathbf{E}_{h-1}^T\mathbf{u}_h/\mathbf{u}_h^T\mathbf{u}_h$
   $\mathbf{t}_h=\mathbf{E}_{h-1}\mathbf{w}_h/\|\mathbf{E}_{h-1}\mathbf{w}_h\|$
   $\mathbf{q}_h=\mathbf{F}_{h-1}^T\mathbf{t}_h/\|\mathbf{F}_{h-1}^T\mathbf{t}_h\|$
   $\mathbf{u}_h=\mathbf{F}_{h-1}\mathbf{q}_h$
4. Calculate the **X**-loadings:
   $\mathbf{p}_h=\mathbf{E}_{h-1}^T\mathbf{t}_h/\mathbf{t}_h^T\mathbf{t}_h=\mathbf{E}_{h-1}^T\mathbf{t}_h$
5. Find the inner model:
   $b_h=\mathbf{u}_h^T\mathbf{t}_h/\mathbf{t}_h^T\mathbf{t}_h=\mathbf{u}_h^T\mathbf{t}_h$
6. Calculate the residuals:
   $\mathbf{E}_h=\mathbf{E}_{h-1}-\mathbf{t}_h\mathbf{p}_h^T$
   $\mathbf{F}_h=\mathbf{F}_{h-1}-b_h\mathbf{t}_h\mathbf{q}_h^T$
7. Return to step 2 until all principal factors are calculated.

1사이의 값을 가지게 되며 경험적으로는 R^2가 Q^2보다 보통 5-20% 정도 더 높게 나타나는데 실제로 X 변수 중 Y와 상관 없는 변수가 많이 포함되어 overfitting이 될수록 이 차이가 더 커지게 된다.

2-2. Recursive PLS Regression

공정 모델은 공정의 시간에 따른 변화를 반영하기 위해 새로운 공정 데이터에 근거해서 갱신 되어야 한다. 모델을 갱신하기 위한 방법인 RPLS 알고리듬이 Helland 등[5]에 의해 제안 되었다. PLS 회귀 모델은 V를 최소화 하기 위해서 $\|Y-XC\|^2$을 최소화하기 위한 C를 다음과 같이 구할 수 있다.

$$C^{PLS}=(X^TX)^+X^TY=W(P^TW)^{-1}BQ^T \tag{6}$$

여기서 W는 X와 Y 사이의 관계를 나타내는 weighting 행렬이다. 새로운 데이터 $\{X_1, Y_1\}$을 이용해서 기존에 있던 PLS 모델을 갱신할 때 사용되는 자료 행렬은 식 (7)과 같다. 식 (7)의 행렬을 가지고 PLS 회귀 계수 행렬을 구하기 위해 식 (6)을 이용하면 식 (8)과 같은 회귀 계수 행렬을 구할 수 있다.

$$X_{new}=\begin{bmatrix}X\\X_1\end{bmatrix} \quad \text{and} \quad Y_{new}=\begin{bmatrix}Y\\Y_1\end{bmatrix} \tag{7}$$

$$C_{new}^{PLS}=\left(\begin{bmatrix}X\\X_1\end{bmatrix}^T\begin{bmatrix}X\\X_1\end{bmatrix}\right)^+\begin{bmatrix}X\\X_1\end{bmatrix}^T\begin{bmatrix}Y\\Y_1\end{bmatrix} \tag{8}$$

T의 열들이 서로 orthonormal 하고 $T^TF=0$ 이므로 다음과 같이 쓸 수 있다[4].

$$X^TX=PT^TTP^T=PP^T \tag{9}$$

$$X^TY=PT^TTBQ^T+PR^TF=PBQ^T \tag{10}$$

식 (8), (9) 그리고 식 (10)에 의해 회귀 계수 행렬은 다음과 같다.

$$C_{new}^{PLS}=\left(\begin{bmatrix}P^T\\X_1\end{bmatrix}^T\begin{bmatrix}P^T\\X_1\end{bmatrix}\right)^+\begin{bmatrix}P^T\\X_1\end{bmatrix}^T\begin{bmatrix}BQ^T\\Y_1\end{bmatrix} \tag{11}$$

식 (8)과 (11)의 비교를 통해, 새로운 데이터를 이용한 RPLS의 모델 갱신은 다음 자료 행렬에 의해 수행된다.

$$X_{new}=\begin{bmatrix}P^T\\X_1\end{bmatrix} \quad \text{and} \quad Y_{new}=\begin{bmatrix}BQ^T\\Y_1\end{bmatrix} \tag{12}$$

즉, PLS로 모델을 갱신할 때는 지난 데이터와 새로운 데이터를 이용하지만, RPLS는 지난 PLS 모델과 새로운 데이터를 이용한다. 그래서 새로 구성된 자료 행렬의 크기가 작기 때문에 PLS 모델을 만드는데 부하가 적고 계산 시간이 단축되는 이점이 있다.

2-3. Block-wise RPLS

RPLS 알고리듬은 몇 개의 새로운 샘플이 들어오면 모델을 갱신한다. 하지만 실제 공정에선 데이터의 상당한 양이 모이거나 공정의 큰 변화가 있을 때까지 모델을 갱신하지 않을 수도 있다. 그래서 몇 개의 샘플보다 훨씬 데이터 수가 많은 블록의 개념을 사용한다[4]. 새로운 데이터 블록이 들어 왔을 때, 기존의 PLS 모델과 새로운 블록을 이용해 만든 PLS 모델을 이용해 모델을 만든다. 새로 만들어진 모델에 의한 자료 행렬이 블록을 압축한 축소 모델이고, 블록 자체는 많은 샘플들로 구성되어 있기 때문에 RPLS보다 모델 갱신을 자주 하지 않아도 된다. 모델에 사용될 자료 행렬은 다음과 같다.

$$X_{new}=\begin{bmatrix}P^T\\P_1^T\end{bmatrix} \quad \text{and} \quad Y_{new}=\begin{bmatrix}BQ^T\\B_1Q_1^T\end{bmatrix} \tag{13}$$

여기서 P^T와 BQ^T는 기존의 PLS 모델에서 나온 행렬이고, P_1^T 과 $B_1Q_1^T$은 새로운 블록의 PLS모델에서 나온 행렬이다. 결국 기존 블록의 모델과 새로운 블록에 대한 모델을 결합한 행렬로 모델을 갱신하는 것이다.

2-4. Block-wise RPLS with a moving window and forgetting factors

모델에 사용된 블록들의 수를 윈도우(window)라 하는데, 이 윈도우 크기를 일정하게 유지하면서 현재 시점의 블록과 지난 블록들의 모델에 forgetting factors를 지수함수 꼴로 감소하면서 곱해 준다. Forgetting factors가 작을수록 지난 데이터를 빨리 잊는다. 일정한 윈도우 크기를 유지하기 위해서 새로운 블록의 모델에 대한 행렬이 들어오면 가장 오래된 블록의 모델에 대한 행렬을 없앤다. 이러한 방법은 온라인 상에서 모델을 갱신하는데 적합하다. 다음은 이 방법에 사용되는 자료 행렬을 나타낸 것이다.

$$\underbrace{\begin{bmatrix}P_S^T\\\lambda P_{S-1}^T\\.\\.\\\lambda^{W-1}P_{S-W+1}^T\end{bmatrix}}_{\text{Input matrix}},\underbrace{\begin{bmatrix}B_SQ_S^T\\\lambda B_{S-1}Q_{S-1}^T\\.\\.\\\lambda^{W-1}B_{S-W+1}Q_{S-W+1}^T\end{bmatrix}}_{\text{Output matrix}} \tag{14}$$

여기서 λ는 forgetting factor, S는 현재의 블록이고 W는 윈도우내 블록들의 수이다. 이 방법은 일정 기간의 데이터만을 이용하므로 부하나 계산속도 측면에서 가장 많은 장점을 가지고 있다.

2-5. 방법론

블록들의 유사성 정보인 상관관계를 반영하는 지수를 결정하는 방법과 이 지수를 모델블록들의 forgetting factor로 적용하는 방법은 다음과 같다.

가. 모델링에 필요한 윈도우 크기를 두 블록으로 유지하면서 새로운 블록을 예측한다. 모델링에 사용된 두 블록 각각의 sub-model을 PLS로 모델링 한다.

나. 두 PLS 모델에 의해 구해진 loading 행렬과 결정계수(R^2)를 이용해 다음과 같이 구한다. 이 식에서 사용된 곱은 element-by-element multiplication이다.

$$P_{1(r_1\times m)}^T\times E-R2Y_{(r_1\times m)}=AA_{(r_1\times m)}=[a_{jk}] \tag{15}$$

$$P_{2(r_2\times m)}^T\times E-R2Y_{(r_2\times m)}=BB_{(r_2\times m)}=[b_{jk}] \tag{16}$$

여기서 1과 2는 모델에 사용된 두 블록, r_1과 r_2는 각 블록의 모델에 대한 LVs의 수, m은 변수의 수, E-R2Y는 주성분별로 표현된 Y에 대한 R^2 열 벡터를 변수 수만큼 확장한 행렬이고($r_i\times m$)은 행렬의 크기를 나타낸다.

다. 식 (15)와 (16)은 각 변수별로 압축된 정보인 다음과 같은 벡터로 변형된다.

$$\left[\sum_{j=1}^{r_1}a_{j1}\sum_{j=1}^{r_1}a_{j2}\sum_{j=1}^{r_1}a_{j3}\cdots\sum_{j=1}^{r_1}a_{jm}\right]=aa_{(1\times m)} \tag{17}$$

$$\left[\sum_{j=1}^{r_2}b_{j1}\sum_{j=1}^{r_2}b_{j2}\sum_{j=1}^{r_2}b_{j3}\cdots\sum_{j=1}^{r_2}b_{jm}\right]=bb_{(1\times m)} \tag{18}$$

라. 식 (17)과 (18)의 elements은 각 블록의 변수 정보를 압축한 값들이다. 이 블록들에 대한 상관성 정보를 얻기 위해 두 벡터의 상관계수(r)를 구한다. 이 상관계수는 모델에 사용된 두 블록의 상관관계를 나타내므로 두 블록으로 모델링 할 때, 이 값을 forgetting factor로 사용하면

된다. 즉, 두 블록 중 과거 블록에 상관계수를 곱하고 현재에 가까운 블록에 1을 곱해서 자료 행렬을 만든 다음, 그 자료 행렬을 이용해서 PLS 모델을 수행한다. 상관계수가 작으면 그만큼 두 블록의 상관성이 작다는 것이므로, 예측하는 블록의 상관성도 떨어질 것이다. 따라서 모델 블록 중 최근 블록에 weight를 더 크게 주고, 상대적으로 과거 블록에 weight를 작게 준다. 두 블록이 상관성이 크다는 것은 그만큼 다음 블록도 상관성이 클 가능성이 높으므로 두 블록사이의 weight 차이를 작게 해서 최대한 정보를 많이 가지고 예측하도록 한다. 이와 같이 상관계수를 forgetting factor로 사용해서 모델을 만드는 것은 모델 블록에 대한 정보로부터 예측 블록에 대한 정보를 최대한 반영하도록 했기 때문에 보다 좋은 예측력을 가진 모델을 만들 수 있다는 장점을 가지고 있다.

3. 사례연구

3-1. 대상 공정

대상 공정은 정유공장의 공정 병목현상 때문에 추가로 신설된 new crude charge heater와 air preheater system으로 crude charge heater unit와 combustion air preheater unit로 구성되어 있다. 위와 같은 시스템은 crude charge heater에서 유출되는 flue gas의 폐열을 회수, 보다 높은 효율을 증진시키기 위하여 flue gas steam상에 고온의 flue gas와 가열로용 연소공기의 열교환을 위한 air preheater를 설치하였으며, 가열로에서의 연소에 필요한 공기는 air stack을 통해서 FDF(forced draft fan)에 의해 공기 예열기에 공급된다. 공기 예열기에서 예열된 hot air는 원유 가열로에서 공기 예열기를 통해 나오는 연소가스 온도가 최저 170 ℃ 이상을 유지하기 위해 필요에 따라 by-pass line을 통해 공기 예열기를 거치지 않고 직접 원유 가열로에 공급되어진다. 원유 가열로에서 연료와 공기는 함께 연소 되어 공정 유체에 열을 전달한 후 나머지 열량이 공기 예열기에서 연소공기와 열교환 되며 IDF(induced draft fan)에 의해 집진 설비인 multi-cyclone을 통해 가열로 stack으로 배출된다. 또한 natural draft 운전의 경우는 FDF와 IDF를 정지시킨 후에 공기 예열기를 거치지 않고 직접 가열로 stack으로 배출된다. 원유 가열로는 연료로서 fuel oil과 fuel gas를 모두 사용할 수 있는 combination type으로 버너에서

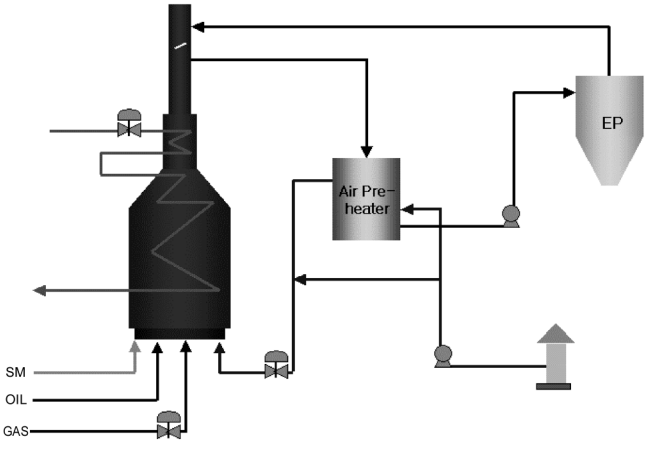


Fig. 1. Schematic diagram of a crude heater and an air preheater.

스팀으로 atomizing 시킨다[13]. 이 공정의 개략도를 Fig. 1에 나타내었다. 공정 변수는 독립 변수 25개와 종속 변수 1개로 구성되어 있다. 독립 변수는 유량, 온도, 압력과 관련된 변수이고, 종속 변수는 온라인 분석기에 의해 측정되는 NO_x 배출 농도 값이다. 샘플은 실시간 데이터베이스(RTDB)로부터 원하는 시간마다 가져올 수 있다. 모델링에 사용된 샘플링 타임은 충분한 데이터를 이용할 수 있고 모델링 결과를 비교적 신뢰할 수 있는 2분을 적용했다. 모델링을 하기 위해 사용된 데이터는 2001년 4월부터 2002년 4월까지의 데이터이고, 11월까지의 데이터는 주로 모델링을 통한 분석 과정에 사용되었고, 11월말부터 2002년 4월까지의 데이터를 제안한 방법으로 모델링하고 테스트하였다.

3-2. NO_x 배출 농도 추정을 위한 모델의 구성

NO_x 배출 농도 추정 모델을 3가지 데이터 그룹들을 통해 구성하였다. 첫 번째 그룹 데이터는 NO_x 배출 농도가 고농도(평균 150-160 ppm)일 때의 데이터이고, 두 번째 그룹 데이터는 저농도(평균 110-120 ppm)일 때의 데이터이다. 이들 두 그룹은 한 블록의 크기를 1주만큼의 샘플수로 하고 새로운 블록이 들어오는 1주일(샘플 수 기준) 단위로 모델을 갱신한다. 즉 과거 두 블록으로 모델을 구성한 다음 현재의 블록을 예측한다. 몇몇 데이터의 NO_x 배출 농도는 같은 그룹의 평균 NO_x 배출 농도보다 훨씬 높거나 낮은 경우도 있다. 실제로 모델을 테스트 하는 과정에서는 이런 데이터들이 많지 않고 PLS모델이 잘 적응해 가므로 크게 영향을 미치지는 않았다. 그리고 세 번째 그룹은 두 번째 그룹 데이터의 일부분이지만 한 블록의 크기를 3일로 해서 같은 방법으로 테스트한 결과이다. PLS 모델링에서 사용된 LVs 개수는 블록마다 다소 차이가 있었지만 3-7개가 사용되었고, R2Y는 75%-95%의 값을 나타내었다.

3-3. 적용 결과

모델 데이터 블록의 정보를 그대로 이용한 λ=1 일 때(block-wise RPLS with a moving window, Qin's method)와 본 저자가 제안한 상관관계를 이용한 λ=r 일 때의 RMSE(root mean square error)값을 비교하였다. 즉, 모델을 만들 때 미리 두 값을 적용하여 예측에 대한 에러값을 구하였다. 그리고 Min. 값은 각 블록 예측 후, 에러가 최소가 되는 λ값에서의 RMSE값을 나타내었다. 즉, Min. 값은 실제값을 예측할 때 나올 수 있는 최소 에러값이고, 모델을 통한 실시간 예측으로는 구할 수 없는 값이다. 그러므로 Min. 값에 가까운 값을 가질수록 더 좋은 예측력을 나타낸다고 할 수 있다. Accuracy는 100×RMSE of Min./RMSE of (λ=r)을 나타내므로 상관계수 값을 forgetting factor로 사용했을 때 얼마나 좋은지를 판단할 수 있는 기준이 된다. 검증에 사용된 모든 데이터는 이상치를 제외한 다음 모델을 만들고 검증하였다.

3-3-1. 2001년11월27일-2002년 1월 8일

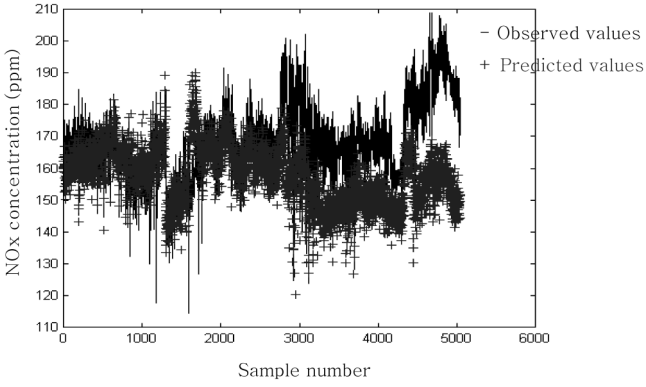
Table 2에 결과를 나타내었고, 사용된 데이터는 5개 블록이다. 첫 번째와 두 번째 블록으로 세 번째 블록을 예측할 때를 보면, λ=r 일 때가 매우 좋은 예측력을 나타내었다. 나머지 블록에 대한 예측에서도 λ=r 일 때가 Min. 값에 근접함을 확인할 수 있다. 세 번째 블록에 대한 결과를 두 λ값에 대해 Fig. 2에 나타내었다. 그림을 통해서 에러값의 차이를 쉽게 확인할 수 있다.

3-3-2. 2002년 2월 5일-2002년 4월10일

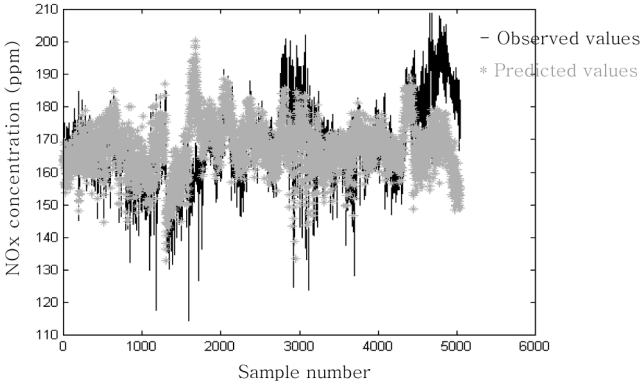
Table 3에 결과를 나타내었고, 사용된 데이터는 8개 블록이다. 대부분

Table 2. Correlation coefficient(r) and RMSE ('01. 11. 27-'02. 1. 8)

Number of prediction blocks	r (model blocks)	λ=1	λ=r	Min.	Accuracy (%)
3(model blocks:1,2)	0.0844	17.5654	11.1224	11.1224	100
4(model blocks:2,3)	0.5158	12.2814	12.9181	11.7056	91
5(model blocks:3,4)	0.1382	12.0109	11.7261	11.4144	97
Total blocks		14.1849	11.9455	11.4166	96

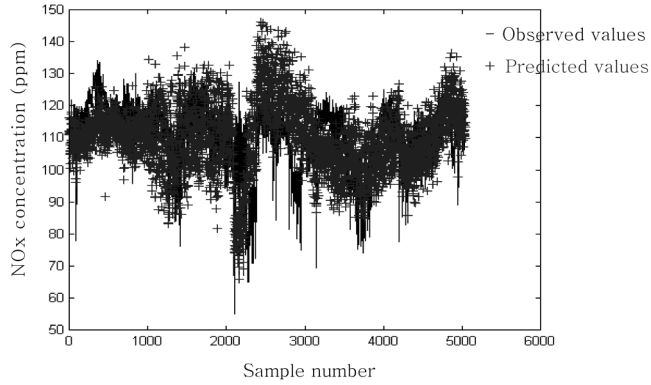


(a) λ=1, rmse=17.5654

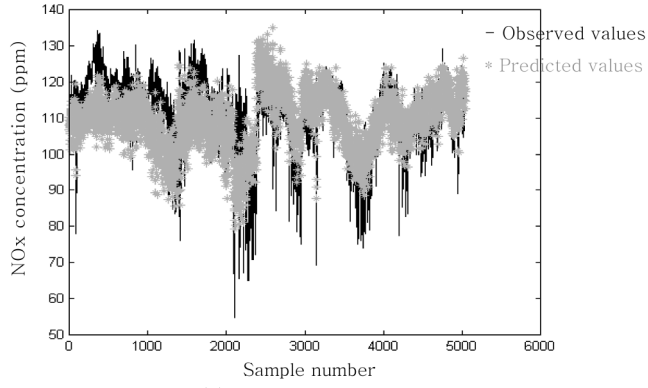


(b) λ=r, rmse=11.1224

Fig. 2. NO_x Concentration for the seventh block in the group I.



(a) λ=1, rmse=12.3775



(b) λ=r, rmse=8.7908

Fig. 3. NO_x Concentration for the seventh block in the group II.

의 블록에서 λ=r 일 때가 λ=1 일 때보다 에러값이 작게 나타났다. 특히 일곱 번째 블록의 예측 에러값은 λ=r일 때가 λ=1 일 때보다 훨씬 작은 값을 나타내었다. 그 결과를 Fig. 3에 나타내었다.

3-3-3. 2002년 2월25일-2002년 3월17일

Table 4에 결과를 나타내었고, 사용된 데이터는 7개 블록이다. 앞선 두 데이터 그룹이 한 블록의 크기를 7일(이상치 제거 후 샘플 수 기준)로 한 것과 달리 이 그룹 데이터에선 한 블록의 크기를 3일로 하여 모델링 하고 검증하였다. 위의 두 가지 경우와 마찬가지로 λ=r 일 때가 에러값이 작고 Min. 값에 가까운 값을 가짐을 확인할 수 있었다.

위의 결과들에서 알 수 있듯이 모델 블록들에 의해 구해진 상관계수 값을 forgetting factor로 하였을 때, 복잡한 최적화 문제의 구성이나 복잡한 수식을 통해 forgetting factor를 구하는 과정이 필요하지 않다. 그리고 예측 블록에 대한 정보를 모르지만 모델 블록에 대한 정보만으로 충분히 좋은 예측력을 나타낼 수 있다는 것을 확인했다. 다시 말해 대상공정의 실제 예측 블록이 모델 블록과 변하는 경향이 비슷할 경우 모델 블록의 상관관계를 이용해 좋은 예측 모델을 만들 수 있었다. Table 2와 3의 가장 두드러진 차이를 보이는 세 번째 예측 블록과 일곱 번째 예측 블록을 보면, λ=r 일 때가 에러가 작은 것을 확인할 수 있다. 이것은 모델의 최근 한 블록과 예측 블록이 모델 블록으로 사용될 때 상관계수 값이 전 모델 블록들의 상관계수 값보다 커질 것을 예상할 수 있다. 결과를 보면 실제로 더 큰 값을 가짐을 볼 수 있다. 그러므로 본 저자가 정의한 상관계수 값이 모델의 상관성을 대표한다고 할 수 있다. 그리고 상관계수인 r은 블록 내에서의 국부 모델인 λ=0과 전역 모델인 λ=1의 모델을 동시에 보완하는 값이라 할 수 있다. 즉, 모델이 국부 모델에 가까우면 r은 0에 가까운 값을, 전역 모델에 가까우면 1에 가까운 값을 나타내었다. 그러므로 모델 블록의 상관성을 나타내는 지표가 중요하고 그 지표를 forgetting factor로 사용해 모델을 만드는 것이 효과적이라고 할 수 있다.

4. 결 론

본 연구에서는 온라인 적응 모델링을 위해 block-wise RPLS with a moving window and forgetting factors 방법을 이용하였고, 상관관계를 쉽게 파악하기 위해 두 개의 블록을 사용하였다. 모델에 사용된 두 블록의 상관성 정보를 얻기 위한 방법을 제시하였고, 그 지표를 forgetting factor로 사용하여 모델링에 사용되는 자료 행렬을 구성함으로써 새로운 샘플들에 대한 예측력을 높일 수 있었다. 그리고 대상 공정에 대한 공

Table 3. Correlation coefficient(r) and RMSE ('02. 2.5-'02. 4. 10)

Number of prediction blocks	r (model blocks)	λ=1	λ=r	Min.	Accuracy (%)
3(model blocks:1,2)	0.5072	8.0975	8.0230	8.0230	100
4(model blocks:2,3)	0.7920	6.8223	6.5220	6.4233	98
5(model blocks:3,4)	0.6914	6.3937	6.5774	6.1595	94
6(model blocks:4,5)	0.6696	8.5597	8.4139	8.0352	95
7(model blocks:5,6)	0.1220	12.3775	8.7908	8.4313	96
8(model blocks:6,7)	0.4189	12.0530	12.5536	12.0530	96
Total blocks		9.3518	8.7162	8.4112	97

Table 4. Correlation coefficient(r) and RMSE ('02. 2. 25-'02. 3. 17)

Number of prediction blocks	r (model blocks)	λ=1	λ=r	Min.	Accuracy (%)
3(model blocks:1,2)	0.7348	4.8875	4.2543	4.0242	95
4(model blocks:2,3)	0.6092	4.6771	4.3217	4.3217	100
5(model blocks:3,4)	0.6832	5.7175	5.6604	4.5403	80
6(model blocks:4,5)	0.3601	6.6184	6.6058	6.4509	98
7(model blocks:5,6)	0.5203	6.1571	6.0845	5.9187	97
Total blocks		5.6598	5.4676	5.1406	94

정 지식 없이 상관성 정보만으로 모델 블록 내에서의 전역 모델과 국부 모델을 동시에 보완하는 모델을 만들 수 있어서 계산 부하가 적고 예측력이 좋은 온라인 적응 모델을 만드는 데 효과적이었다.

감 사

본 연구는 두뇌한국21 사업의 화공 사업단인 포항공과대학교 화학공학과와 교육부 지정국책대학원인 포항공과대학교 환경공학부의 지원으로 이루어진 것으로 이에 감사를 드립니다.

사용기호

AA : $\in R^{r1\times n}$, matrix including information in sub-block I
a_{jk} : elements of **AA** matrix
aa : row vector including compressed information of variables in AA
B : diagonal matrix of inner model coefficients in X and Y
B_1 : diagonal matrix of inner model coefficients in X_1 and Y_1
BB : $\in R^{r2\times n}$, matrix including information of sub-block II
b_{jk} : elements of BB matrix
bb : row vector including compressed information of variables in BB
C : model coefficient matrix
C^{PLS} : regression coefficient matrix from PLS
C^{PLS}_{new} : updated regression coefficient matrix from PLS
E : residual matrix for X
F : residual matrix for Y
G : residual matrix for U
m : number of variables in X
P : loading matrix for X
P_1 : loading matrix for X_1
Q : loading matrix for Y
Q_1 : loading matrix for Y_1
Q^2 : cross-validated coefficient of determination
R^2 : coefficient of determination
r : correlation coefficient between model blocks
r_1 : number of latent variables in PLS model of sub-block I
r_2 : number of latent variables in PLS model of sub-block II
S : current block
T : score matrix for X
U : score matrix for Y
V : noise matrix
W : weighting matrix in PLS
W : number of block in a moving window
X : input data matrix
X_1 : new input data matrix
X_{new} : updated input data matrix
Y : output data matrix
Y_1 : new output data matrix
Y_{new} : updated output data matrix

그리이스문자

λ : forgetting factor

윗첨자

T : transpose

참고문헌

1. Nomikos, P. and MacGregor, J. F., "Multivariate SPC Charts for Monitoring Batch Processes," *Technometrics*, **37**, 41-59(1995).
2. Wise, B. M. and Gallagher, N. B., "The Process Chemometrics Approach to Process Monitoring and Fault Detection," *J. Proc. Cont.*, **6**(6), 329-348(1996).
3. MacGregor, J. F., Jaeckle, C., Kiparissides, C. and Koutoudi, M., "Process Monitoring and Diagnosis by Multiblock PLS Methods," *AIChE J.*, **40**(5), 826-838(1994).
4. Qin, S. J., "Recursive PLS Algorithms for Adaptive Data Modeling," *Computers Chem. Engng.*, **22**(4/5), 503-514(1998).
5. Helland, K., Berntsen, H. E., Borgen, O. S. and Martens, H., "Recursive Algorithm for Partial Least Squares Regression," *Chem. and Int. Lab. Sys.*, **14**, 129-137(1992).
6. Dayal, B. S. and MacGregor, J. F., "Recursive Exponentially Weighted PLS and Its Applications to Adaptive Control and Prediction," *J. Proc. Cont.*, **7**(3), 169-179(1997).
7. Chakravarthy, S. S. S., Vohra, A. K. and Gill, B. S., "Predictive Emission Monitors (PEMS) for NO_x Generation in Process Heaters," *Comp. Chem. Eng.*, **23**, 1649-1659(2000).
8. Collins, M. and Terhune, K., "A Model Solution for Tracking Pollution," *Control Engineering*, June(1994).
9. Faravelli, T., Bua, L., Frassoldati, A., Antifora, A., Tognotti L. and Ranzi, E., "A New Procedure for Predicting NO_x Emissions from Furnaces," *Comp. Chem. Eng.*, **25**, 613-618(2001).
10. Kocijan, J., "An Approach to Multivariate Combustion Control Design," *J. of Process Control*, **7**(4), 291-301(1997).
11. Geladi, P. and Kowalski, B. R., "Partial Least-Squares Regression: A Tutorial," *Anal. Chim. Acta.*, **185**, 1-17(1986).
12. Wold, S., "Cross Validatory Estimation of the Number of Components in Factor and Principal Component Analysis," *Technometrics*, **20**, 397-406(1978).
13. Borman, G. L. and Ragland, K. W., *Combustion Engineering*, Mc Graw Hill, Boston(1998).