

## PLS 방법을 이용한 화학공정 품질변수의 적응 예측 모델 개발

김성영, 이범석\*  
 경희대학교 화학공학과  
 (bslee@khu.ac.kr\*)

**Development of the Adaptive Predictive Model for product qualities in chemical processes using PLS method**

Sung Young Kim, Bomsock Lee\*  
 Department of Chemical Engineering, Kyung Hee University  
 (bslee@khu.ac.kr\*)

**서론**

본 논문에서는 부분 최소 자승법(partial least squares, PLS)을 이용하여 실시간으로 얻어지는 화학공정 데이터에 적응하는 예측 모델을 수립하였다. PLS는 품질변수와 품질에 영향을 미치는 가능한 모든 변수들과의 상관관계를 설정하는 방법으로 공정으로부터 얻어지는 데이터를 이용하여 모델을 구성하는데 있어 매우 유용하다. 이러한 PLS 방법을 개선하여 기존의 공정 데이터를 이용하여 예측 모델을 만들고 새로운 데이터가 들어오면 기존 예측 모델에서 예측이 좋지 않은 데이터를 제거하고 새로운 데이터를 모델에 포함시키는 방법을 이용하여 적응 예측 모델을 만들었다. NIR Diesel Fuel Spectra 데이터에 이 적응 예측 모델을 적용하여 기존의 방법 중 업데이트 하지 않은 PLS, 실시간으로 들어오는 데이터를 모두 포함시켜서 실행한 PLS, Recursive PLS 방법과 예측 성능을 비교해 보았다.

**이론****Partial Least Squares**

PLS 방법은 새로운 잠재 변수(latent variable)들을 찾아내 모델에 사용하는 방법으로 PLS에 사용되는 잠재변수들은 서로 선형적으로 독립이며 입력 잠재변수와 출력 잠재변수 사이에는 매우 높은 상관관계를 갖게 되며 입력 변수 중에 분산(variance)값이 높은 순서대로 잠재변수로 채택된다.

입력 변수 데이터 블록( $X$ )은 score vector  $t_h(h=1, \dots, a)$ 와 load vector  $p_h(h=1, \dots, a)$ 의 곱으로 나타내어지고, 출력 변수 데이터 블록( $Y$ )은 score vector  $u_h(h=1, \dots, a)$ 와 load vector  $q_h(h=1, \dots, a)$ 의 곱으로 표시되어진다.

$$X = \sum_{h=1}^a t_h p_h^T + E \quad (1)$$

$$Y = \sum_{h=1}^a u_h q_h^T + F \quad (2)$$

$$u_h = \sum_{h=1}^a b_h t_h + r_h \quad (3)$$

PLS를 수행하는 대표적인 NIPALS 알고리즘을 이용하여  $t_1$ 과  $u_1$ 이 구하여 다음 식으로 두 score vector간의 회귀계수를 구한다.

$$b_1 = \frac{u_1^T t_1}{t_1^T t_1} \quad (4)$$

계산된 load vector  $p_h, q_h (h=1, \dots, a)$ 와 회귀계수  $b_h(h=1, \dots, a)$ 를 사용하여 학습에 사용되었

거나 혹은 사용되지 않은 입력변수 데이터 블록  $X$ 에 대한 출력변수 데이터블록  $Y$ 의 예측값을 계산한다.

### Recursive PLS

PLS 방법을 이용하여 새로운 데이터  $X_1, Y_1$ 가 얻어지면 이를 이용해서 기존의 PLS모형을 업데이트 하는 방법이다.

먼저 PLS를 수행하고,

$$[X, Y] \rightarrow \text{PLS} \rightarrow [T, W, P, B, Q] \quad (5)$$

새로운 데이터  $\{X_1, Y_1\}$ 를 이용하여 다음과 같이 RPLS 모델 업데이트가 수행된다.

$$X_{new} = \begin{bmatrix} P^T \\ X_1 \end{bmatrix}, \quad Y_{new} = \begin{bmatrix} BQ^T \\ Y \end{bmatrix} \quad (6)$$

### Adaptive Predictive PLS

이 방법은 기존의 PLS 방법을 개선하여 실시간으로 얻어지는 공정 데이터를 모델에 업데이트 하는 방법이다.

위의 RPLS 방법과 달리 Adaptive Predictive PLS는 새로운 데이터  $X_1, Y_1$ 가 얻어지면 기존의 PLS 모델에 사용되었던 데이터 중에서 예측이 가장 낮은 데이터를 제거하고 새로운  $\{X_1, Y_1\}$ 를 추가하여 PLS 모델을 업데이트 시키는 방법이다. 즉, PLS 모델에 사용되는 데이터의 수는 일정하게 유지되고 동시에 모델에 포함된 데이터 중 예측에 나쁜 영향을 미치는 데이터는 계속 제거되어 모델의 예측성능을 향상시켜주는 이점이 있다.

### 사례연구

NIR Diesel Fuel Spectra 데이터를 이용하여 위의 방법들을 적용시켜 보았다. near IR을 이용한 실시간 모니터링을 위한 목적으로 사용되었고, 401 wavelengths-wide NIR spectra를 입력 값으로 사용하여 출력 값으로 cetane number를 예측해 보았다. 94 samples를 이용하여 모델을 수립하고 18 samples를 이용하여 실시간으로 예측하고 업데이트하여 예측성능을 비교해보았다.

적용은 아래의 4가지 방법을 이용하였다.

- ① PLS : 94 samples로 PLS 모델을 만들어 18 samples를 업데이트 없이 예측.
  - ② PLS : 94 samples로 PLS 모델을 만들고 실시간으로 업데이트되는 18 samples를 모델에 하나씩 추가하여 예측.
  - ③ RPLS
  - ④ Adaptive Predictive PLS
- ②,③,④에서는 모델을 만들고 예측 블록인 18 samples 중 1 sample을 예측한 후, 그 sample의 실제 공정 값이 얻어지면 sample을 모델에 업데이트하고 그 다음 sample을 예측하는 방법으로 진행하였다.

위의 4가지 방법들을 적용하여 실제 공정 값과 예측 값의 Error를 구하고, 그 Error값을 이용해서 예측성능을 비교하였다.

### 결과 및 토론

그림 1.은 위의 4가지 방법을 이용하여 예측한 값과 실제 공정 값과의 Root Square Error를 나타낸 것이다.

첫 번째는 94 samples로 PLS 모델을 만들고 실시간으로 들어오는 데이터를 업데이트 하지 않은 상태에서 각각의 모르는 출력 값을 예측한 결과이다. 이 방법은 모델에 사용되는 데이터

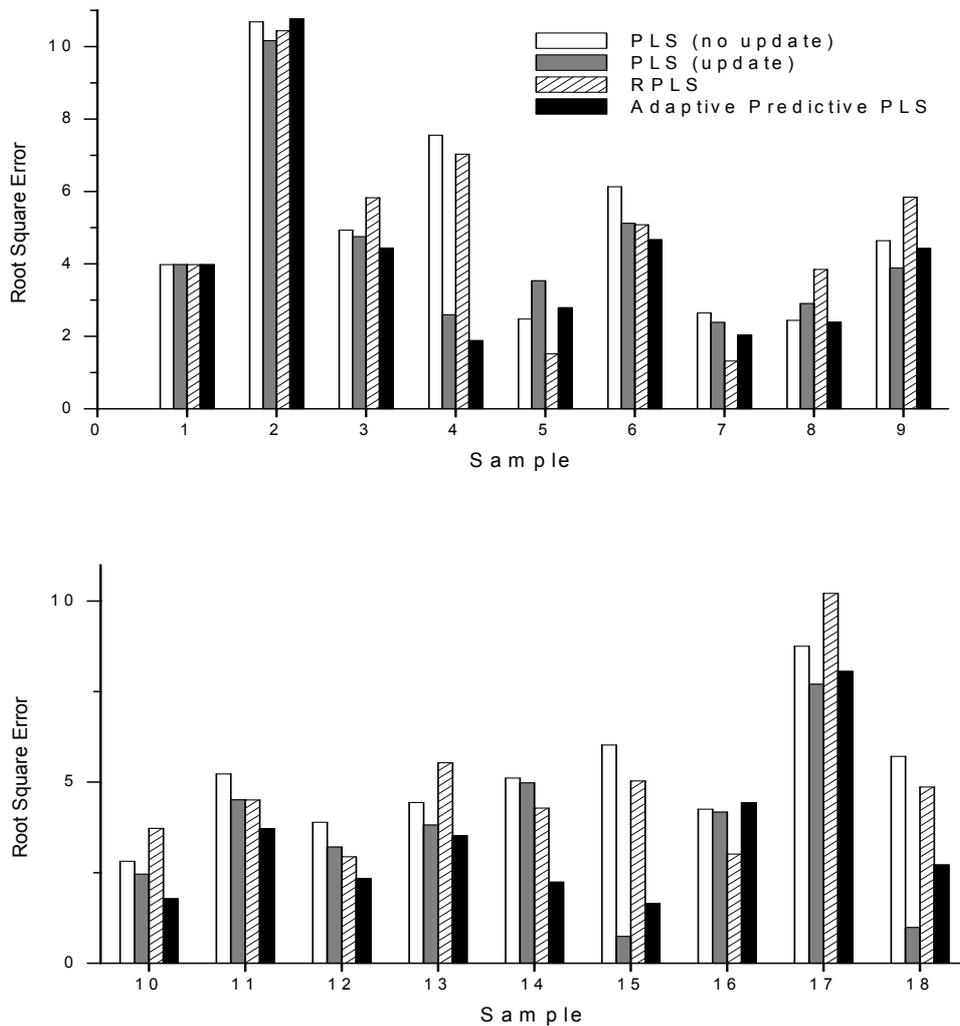


그림 1. Root Square Error for Prediction

의 수가 증가하지 않고 출력 값의 계산이 용이한 이점이 있지만 기존의 Error가 계속 모델에 포함되고 공정의 변화에 대응하지 못하여 시간이 많이 흘러 공정의 변화가 생기면 출력 값의 Error가 커지게 된다.

두 번째는 PLS 모델을 이용하여 실시간으로 들어오는 18 samples을 모델에 업데이트 시켜서 예측한 결과이다. 이 방법은 새로운 데이터가 추가 될 때 마다 새롭게 PLS를 수행하므로 다음 출력 값의 예측 성능이 좋아진다. 그러나 오랜 시간이 지나 데이터의 량이 많아지면 모델에 사용되는 데이터의 수가 무한히 증가하게 되어 모델을 수립하기 위한 계산에 많은 시간이 필요하게 된다.

세 번째는 RPLS 방법을 이용하여 새로운 데이터를 실시간으로 업데이트 하면서 다음 출력 값을 예측한 결과이다. 업데이트를 하지 않은 첫 번째 PLS보다는 예측 성능이 향상되지만 기존의 PLS 모델과 새로운 데이터를 이용하여 모델을 업데이트 하는 방법이므로 기존의 모델에 포함된 Error가 누적되는 현상이 발생하게 된다.

네 번째는 본 연구에서 제안한 Adaptive Predictive PLS 방법을 이용하여 데이터를 예측하고 업데이트 한 결과이다. 이 방법은 RPLS보다는 계산에 조금 더 많은 시간이 필요하지만 새로운 데이터를 계속 누적시키는 두 번째 PLS 방법과는 달리 모델에 사용되는 데이터의 수가

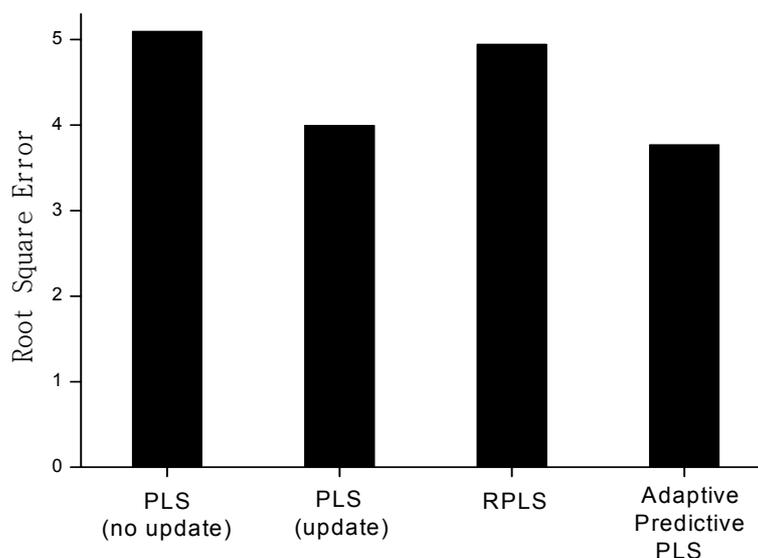


그림 2. Root Square Error의 평균.

증가하지 않으므로 계산 시간이 일정하게 유지된다. 그리고 새로운 데이터가 얻어짐과 동시에 기존의 예측이 좋지 않은 데이터는 모델에서 제외되게 되므로 시간이 흐름에 따라 모델에 포함된 Error도 제거되며 공정의 변화에도 잘 적응하는 결과가 나타나게 된다.

각각의 방법에 대한 Root square error의 평균을 구하여 그림 2.로 나타내었다. 다른 방법들과 비교하여 Adaptive Predictive PLS의 예측 성능이 뛰어난 것을 확인할 수 있다. 이와 같이 Adaptive Predictive PLS는 기존의 방법들보다 실시간으로 변화하는 화학공정의 출력 값을 예측하는데 유용하다는 것을 알 수 있다.

#### 감사의 글

본 연구는 한국과학재단(과제번호 R01-2003-000-10697-0)에서 지원되었으며 이에 감사를 드립니다.

#### 참고문헌

1. S. Joe Qin, "Recursive PLS algorithms for adaptive data modeling", *Computers & Chemical Engineering*, vol. 22, no. 4/5, 503-514, 1998.
2. Paul Geladi and Bruce R. Kowalski, "Partial Least-Squares Regression : A tutorial", *Analytica Chimica Acta*, 185, 1-17, 1986.
3. Zvi Boger, "Selection of quasi-optimal inputs in chemometrics modeling by artificial neural network analysis", *Analytica Chimica Acta*, 490, 31-40, 2003.