비선형 부분최소자승법을 이용한 가스공정의 추론모델 개발

손정현, 송지호, 신동일*, 윤인섭
서울대학교 응용화학부, *서울대학교 화학공정신기술연구소

Inferential Model Development for Gas Process Using Nonlinear Partial Least Square

Jung Hyun Sohn, Ji Ho Song, Dongil Shin*, En Sup Yoon
School of Chemical Engineering, Seoul National University, Seoul 151-742, Korea
**Institute of Chemical Processes, Seoul National University, Seoul, 151-742, Korea*

1. 서 론

물질의 조성(concentration)이나 분자량과 같은 변수는 생산물의 품질이나 효과적인 공정조업과 직결되는 변수로써 실시간으로 측정되어야 한다. 그러나 GC(gas chromatography)와 같은 실시간 분석기는 가격이 비싸고, 채취(sampling)가 제한되어 있으며 채취 및 분석시간이 길어서 측정신뢰도가 떨어지고 비싼 유지, 보수비용을 필요로 한다. 추론모델(inferential model)은 이와 같이 높은 압력, 온도 하에서의 조업, 비싼 설치, 유지, 보수비용 등의 이유로 센서의 설치가 불가능하거나, 측정지연 등으로 실시간 측정이 어려운 경우 신속하고 쉽게 실시간 측정이 가능한 변수로부터 측정이 힘든 변수를 예측하는 모델이다.

PLS(partial least square or projection to latent structure)는 노이즈가 심하고 변수간의 상관관계가 강하며, 제한된 수의 데이터만이 존재하는 문제들에 대해 강력한 회귀성능을 보이는 다변량통계분석법으로 외적변수변환과 내적회귀모델로 구성된다. 선형 PLS는 외적변환과 내적회귀모델로 선형방법들을 사용하게 되는데, 실재적으로 가스, 화학공정의 데이터와 같은 실재데이터에는 비선형 관계가 존재하는 것이 일반적이므로 이러한 비선형 관계성을 설명하기 위해서는 선형모델로는 불충분하며 새로운 비선형 모델이 필요하게 된다.

본 연구에서는 선형 PLS를 개선하여 외적변수변환은 사영기반 알고리즘인 Principal Curve 알고리즘을 신경망으로 구현한 SOFM(self-organizing feature map)을 이용하였고, 내적회귀모델은 다층신경회로망(multilayer feedforward neural network)을 사용하는 비선형 PLS를 제안하였다. 그리고, 제안된 비선형 PLS를 함수데이터와 증류탑의 탑상물질의 조성을 예측하는 추론모델을 수립하는데 적용하였다.

2. 이 론

1. PLS(partial least square)

PLS는 데이터의 노이즈가 심하고 변수간의 상관관계가 강한 경우에 강력한 회귀성능을 보이는 다변량 통계분석 기법으로 외적관계(outer relation)와 내적관계(inner relation)로 이루어지며 MLR(multiple linear regression)이나 PCR(principal component regression)에 비해 뛰어난 강건성(robustness)을 보인다. 외적관계는 변수변환으로 PCA 구현을 위한 NIPALS 알고리즘을 이용하여 다변량 입출력 데이터가 각각 선형변환(linear transformation)을 거쳐 입출력 score를 만드는 과정이다. 내적관계는 회귀모델로 최소자승법(least square)을 이용하여 외적변환을 거쳐 나온 입출력 score 벡터들간의 선형회귀모델을 수립한다. PLS의 내적, 외적 관계를 식으로 나타내면 다음과 같다.

1) 외적관계

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

$$Y = UQ^T + F = \sum_{h=1}^{a} u_h q_h^T + F$$

2) 내적관계

$$\hat{u}_h = b_h t_h \ (h = 1, 2, \cdots, a)$$

$$b_h = u_h^T t_h / t_h^T t_h$$

여기서 X($R^{n\times m}$), Y($R^{n\times q}$)는 입출력 데이터, T($R^{n\times a}$), U($R^{n\times s}$)는 입출력 score vector, P($R^{m\times a}$), Q($R^{q\times a}$)는 입출력 loading vector, E($R^{n\times m}$), F($R^{n\times q}$)는 에러, b는 선형회귀계수를 의미한다.

2. NNPLS(neural network PLS)

데이터에 존재하는 비선형성을 설명하기 위해 McAvoy(1992)는 한 개의 은닉층(hidden layer)을 가지며 활성함수로 시그모이드 함수를 사용하는 다층신경망을 입출력 score vector의 비선형 회귀모델로 사용하는 NNPLS를 제안하였다.

NNPLS에서 내적관계는 다음식으로 나타내지며, N( )은 신경망으로 표현되는 비선형 관계를 의미한다.

$$\hat{u} = N(t_h) + r_h$$

내적회귀모델로 신경망을 사용하는 것은 신경망의 뛰어난 비선형근사성질을 활용하는 것이 되는데, 임의의 연속함수는 신경망을 통하여 원하는 정확도로 근사가 가능하다.

3. 자동연상신경망을 이용한 비선형 PLS

Malthouse 등은 Krammer의 자동연상신경망이 비선형 PCA의 역할을 한다는

것을 이용하여 PLS의 외적변환에 자동연상신경망을 이용한 비선형 PLS를 제안하였다. Krammer의 자동연상신경망은 3개의 은닉층을 가지며, 입력과 출력이 같은 신경망으로 학습 후 주어진 입력과 같은 출력을 낸다. mapping/demapping 층에는 시그모이드 활성함수를 나머지 층에는 선형활성함수를 사용한다. 병목층의 노드수는 입출력 노드수보다 반드시 작아야 하며 이를 통해 정보 압축이 일어나고 병목층의 값은 비선형 PC의 score 값으로 생각할 수 있다.

그러나 자동연상신경망은 3개의 은닉층을 가지며 각각을 독립적으로 학습시킬 수 없기 때문에 일반적인 오류역전파(error back-propagation) 알고리즘을 이용하는 경우 신경망 학습이 어렵고 상당한 시간이 요구되며, 은닉층이 한 개인 경우 cross validation을 통해 최적의 은닉노드수를 결정할 수 있지만 3개의 은닉층이 독립적이지 않은 경우 최적의 은닉층 노드수를 구하는 것이 쉽지 않다. 또, 병목층의 값이 비선형 PC의 score를 나타낸다고 하지만, Edward C. Malthouse는 Krammer 자동연상신경망의 병목층 값이 엄격한 의미에서 score가 아님을 증명하였다. 이는 자동연상신경망에 의한 비선형 특성변수추출(feature extraction)이 연속함수로의 사영으로 제한되기 때문이다.

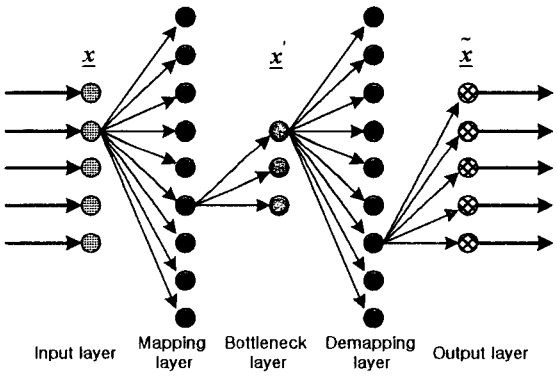

그림 1. Krammer's autoassociative neural network

4. SOFM을 이용한 비선형 PLS

본 연구에서는 PLS의 외적변수변환으로 Principal Curve 알고리즘의 신경망적 구현인 SOFM을 이용하고 내적회귀모델로 시그모이드 활성함수의 은닉층을 가지는 3층 신경망을 사용하였다. Principal Curve는 Hastie와 Stuetzle(1989)에 의해 제안된 방법으로 첫 번째 선형PC의 비선형 일반화라고 할 수 있으며 다음의 self-consistency principle을 만족해야 한다.

$$\hat{x} = F(G(x)) = E(x \mid z = \underset{z}{\arg} \min ||F(z) - x||^2)$$

이는 사영(projection)단계와 조건부평균단계로 나눌 수 있는데, 사영단계에서는 모든 데이터를 곡선 위로 직교사영 시키게 되며, 조건부평균단계에서는 곡선

상의 동일한 위치로 사영되어 같은 z값을 가지는 모든 데이터의 평균값을 z에 해당하는 곡선상의 값으로 한다. 그러나, 실재의 경우에는 유한한 데이터를 다루므로 다수의 점이 principal curve의 동일 지점으로 사영되는 경우가 드물다. 그러므로, 곡선상의 점에서의 평균은 국부가중회귀나 커널회귀 등을 이용하여 span이나 커널폭으로 정의되는 인접영역내의 국부점에 대해서도 가중평균을 수행하여 구하게 된다.

SOFM에서는 입력데이터의 공간분포를 가장 잘 설명하도록 b개의 중심을 위치시키고 입력공간내의 모든 점을 이 b개의 중심으로 사영시킨다. 즉, b개의 중심은 principal curve의 모양을 결정하기도 하고 입력데이터의 특성변수값을 의미하기도 한다. 이러한 성질은 중심개수(b) 결정에 있어 trade-off 문제를 발생시키는데 b의 수를 많게 하면 입력데이터들이 다양한 특성변수값을 가지게 되지만 과적합 곡선이 얻어질 수 있고, b의 수를 작게 하면 과적합의 위험은 없지만 입력데이터들이 다양한 특성변수값을 가지지 못하므로 데이터간의 구분이 분명치 않아진다. PLS에서와 같이 SOFM을 통해 얻어진 score 간의 내적회귀를 수행해야 하는 경우에는 입출력 score 간에 일대일 대응이 이루어지지 않으므로 학습이 잘 이루어지지 않는다. 따라서, 본 연구에서는 이를 극복하기 위해 우선 b개의 중심을 이용하여 입력데이터의 분포를 설명하는 principal curve를 구하고, curve가 구해진 후에는 spline을 이용하여 b개의 각 중심들 사이에 추가적으로 k개의 중심을 생성하고 이 b+k 개의 중심위로 데이터를 사영시켰다.

3. 사례연구

함수데이터와 증류탑의 탑상물질의 조성을 예측을 위한 추론모델 수립에 제안된 비선형 PLS를 이용하였고, 기존의 PLS와 비교하였다. 각 기법의 성능 비교는 테스트 데이터에 대한 mean square error of prediction을 통하여 이루어졌다.

$$MSEP = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

1. 함수데이터

SOFM을 이용한 비선형 PLS 기법을 다음과 같이 한 개의 특성변수에 의해 지배되는 세 개의 입력변수를 가지는 함수데이터에 적용하였다. t∈[1,2]이며 100개의 학습데이터는 N(0, 0.01)의 노이즈를 가진다. 테스트를 위해 역시 N(0, 0.01)의 노이즈를 가지는 100개의 데이터를 얻었다.

$$x_1 = t + e_1$$
$$x_2 = t^2 - 3t + e_2$$
$$x_3 = -t^3 + 3t^2 + e_3$$
$$y = (t-1.25)^2(t-1.9)$$

표 1 과 그림 2에서 알 수 있듯이 제안된 SOFM을 이용한 비선형 PLS가 선형 PLS나 NNPLS보다는 모델차원, 예측능력의 면에서 모두 뛰어난 것을 알 수 있다. 또, 자동연상신경망을 이용한 비선형 PLS와 비교할 때 비슷한 예측성능을 보인다. 표 3에서 PC의 개수는 가장 작은 MSEP를 나타낼 때의 principal component(선형 PLS, NNPLS) 또는 principal curve의 수를 말한다.

표 1. Performance of PLS's

	선형 PLS	NNPLS	AANN PLS	SOFM PLS
# of PC	3	2	1	1
MSEP	0.1056	0.0486	0.0013	0.0015
MSEP(PC=1)	0.7357	0.1680	0.0013	0.0015

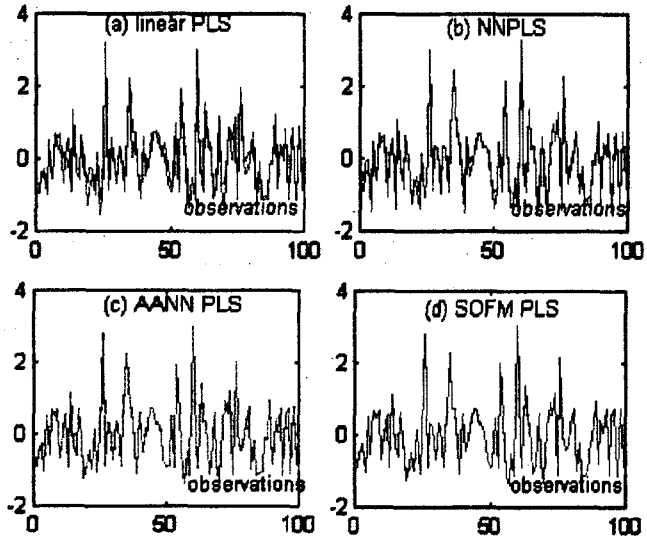

그림 2. PLS prediction performance

2. 증류공정 추론모델

적용을 위해 HYSYS를 이용하여 물과 메탄올 분리 증류탑을 대상으로 100개의 데이터를 얻었다. 증류탑은 15단이고 8단으로 feed가 공급된다. 100개 데이터 중 75개는 학습에 25개는 테스트에 쓰였다. 입력변수로는 1단에서 8단까지의 온도와 환류비를 사용하였고 증류탑 탑상 물질의 조성이 출력변수가 된다.

각단의 온도는 다음식을 이용하여 무차원화 되었는데, θ_L은 가벼운 물질의 특성온도로 condenser의 온도를 이용하였고 θ_H는 무거운 물질의 특성온도로 물의 끓는점을 사용하였다.

$$L = \ln\left(\frac{\theta - \theta_L}{\theta_H - \theta}\right)$$

1개의 PC를 사용한 경우에 대해 PLS의 예측성능을 비교한 결과는 표 2와 같다. 역시, SOFM을 이용한 PLS는 선형 PLS, NNPLS보다 낮은 MSEP값을 가지며, 자동연상신경망을 이용한 비선형 PLS와는 비슷한 예측성능을 보이고 있다.

표 2. Performance of PLS's

	선형 PLS	NNPLS	AANN PLS	SOFM PLS
MSEP(PC=1)	0.1022	0.0562	0.0285	0.0289

4. 결 론

본 연구에서는 비선형성을 가지는 시스템을 모델링하기 위해 spline을 사용하는 개선된 SOFM을 PLS 외적변수변환에 사용하고 다층신경망을 내적회귀모델로 사용하는 비선형 PLS를 제안하였다. 제안된 비선형 PLS는 선형 PLS나 NNPLS에 비해 모델차원측면이나 예측성능면에서 뛰어난 결과를 나타내었고, 자동연상신경망을 외적변수변환으로 사용하는 AANN PLS와 비교할 때 비슷한 예측성능을 나타내었다. 그러나, AANN은 3층의 은닉층으로 인하여 모델의 크기가 커서 많은 학습데이터를 요구하며 모델의 최적 구조 결정이 어렵고 학습에 많은 시간이 걸리는 반면 SOFM을 사용하는 경우 중심개수 b의 조정을 통해 효과적으로 모델의 구조를 결정할 수 있고, 학습에도 많은 시간이 소요되지 않는다.

감 사

본 연구는 교육부의 두뇌한국21사업의 지원에 의해 이루어졌으므로 이에 감사드립니다.

참 고 문 헌

[1] G. Baffi, E. B. Martin, A. J. Morris, "Non-linear projection to latent structures revisited(the neural network PLS algorithm)", Comp. & Chem. Eng., 23, pp. 1293-1307, 1999.

[2] S. J. Qin and T. J. McAvoy, "Nonlinear PLS Modeling Using Neural Networks", Comp. & Chem. Eng., 16(4), pp 379-391, 1992.

[3] E. C. Malthouse, A. C. Tamhane and R. S. H. Mah, "Nonlinear Partial Least Squares", Comp. & Chem. Eng., 21(8), pp. 875-890, 1997.

[4] Paul Geladi, Bruce R. Kowalski, "Partial Least Squares Regression: A Tutorial", Analytica Chimica Acta, 185, pp 1-17, 1986.

[5] Mark A. Kramer, "Nonlinear Principal Component Analysis Using Autoassociative Neural Networks", AIChe Journal, 37(2), pp 233-243, 1991

[6] 정신호, "비선형 주성분 해석을 이용한 화학공정의 이상감지에 관한 연구", 석사학위논문, 서울대학교 응용화학부, 2000.