비선형 주성분해석과 신경망에 기반한 비선형 PLS

Non-linear PLS based on non-linear principal component analysis and neural network

°손 정 현*, 정 신 호**, 송 상 옥***, 윤 인 섭****

* 서울대학교 응용화학부(Tel : 82-02-873-2605 ; Fax : 82-02-884-0530 ; E-mail : jhsohn@pslab.snu.ac.kr)
** 서울대학교 응용화학부(Tel : 82-02-873-2605 ; Fax : 82-02-884-0530 ; E-mail : shinhoj@pslab.snu.ac.kr)
*** 서울대학교 응용화학부(Tel : 82-02-873-2605 ; Fax : 82-02-884-0530 ; E-mail : andy@pslab.snu.ac.kr)
**** 서울대학교 응용화학부(Tel : 82-02-873-2605 ; Fax : 82-02-884-0530 ; E-mail : esyoon@pslab.snu.ac.kr)

Abstract : This paper proposes a new nonlinear partial least square method that extends the linear PLS. Proposed nonlinear PLS uses self-organizing feature map as PLS outer relation and multilayer neural network as PLS inner regression method.

Keywords : non-linear PLS, non-linear PCA, self-organizing feature map, neural network

1. 서론

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 화학공정의 데이터와 같은 실재데이터에는 비선형 관계가 존재하는 것이 일반적이므로 이러한 비선형 관계성을 설명하기 위해서는 선형모델로는 불충분하며 새로운 비선형 모델이 필요하게 된다. 이를 위해 내적회귀모델로 2차회귀모델을 이용하는 QPLS(quadratic PLS, Wold, 1989), 시그모이드 활성함수 신경망을 이용한 NNPLS(neural network PLS, McAvoy, 1993), RBF(radial basis function) 신경망을 이용한 RBFPLS(Wilson, 1997) 등이 제안되었다. 이 방법들은 외적변수변환에는 선형모델을 이용하고 있으며 내적회귀모델로 여러 가지 종류의 비선형 회귀모델을 사용하였다. E. Malthouse(1997)는 외적변수변환을 위해 자동연상신경망(AANN)을 사용하고 내적회귀모델로는 신경망을 사용하는 비선형 PLS를 제안하였다.

화학공정은 측정변수가 많고 변수간 상호연관성이 크기 때문에 PLS, PCA(principal component analysis)와 같은 다변량 통계기법들이 공정감시, 공정모니터링, 추론모델개발, 공정제어 등에 사용되었다. 이러한 방법들은 여러 물리법칙(physical law)에 기반한 수학적인 모델을 세울 수 없거나 수학적 모델을 세우는 것이 가능할 지라도 화학공정과 같이 수많은 재환류와 제어를 위한 복잡한 루프, 공정이상, 외란 등으로 인하여 이론적인 모델로 공정거동을 설명하는 것이 어려운 경우, 과거의 공정데이터를 이용하여 실험적 모델(empirical model)을 세우는데 사용되고 있다. 실험적 모델은 데이터를 이용하여 모델의 구조와 모델 변수를 결정하는 학습과정(training)과 학습하지 않은 데이터에 대한 모델의 일반화 능력을 테스트하는 과정을 거쳐 수립된다. 실험적 모델은 데이터를 기반으로 하므로 모델수립을 위한 데이터는 공정의 여러 가지 거동을 포함하고 있어야 하며 이상거동을 보이는 데이터(outlier)는 모델수립 이전에 제거되어야 한다.

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

2. 이론

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.

PCA는 데이터 공간에서 데이터분포가 넓은 축부터 차례대로 서로 직교하도록 새로운 축들을 정의하고 이들 축으로 정사영 되는 값을 그 축에 대한 새로운 좌표값으로 하는 변환이다. 이때 새롭게 정의된 축을 PC(principal component)라고 하며 축의 방향벡터를 loading vector, 축으로의 정사영 값을 score라고 한다. PCA 이론에 의하면 원형변수의 loading vector는 자료행렬 $X(R^{n\times m})$의 공분산행렬(covariance matrix)의 고유벡터(eigenvector)가 된다. 고유값(eigen value) $\lambda_i(i=1,2,\cdots,m)$를 크기순으로 배열할 때, λ_i에 해당하는 고유벡터가 i번째 PC의 loading vector가 된다. 이 과정에서 작은 고유값을 가지는 PC를 제거하게 되면 데이터의 정보를 잃지 않으면서 원형변수(original variable)보다 작은수의 특성변수(feature variable)를 구할 수 있게 되어 데이터 차원감소를 이룰 수 있게 된