Chemometric 방법을 이용한 Gross Error Detection

이은룡, 조병학, 김인원
건국대학교 화학공학과

Gross Error Detection using Chemometric Methods

Eun-Lyong Lee, Byoung-Hak Cho and In-Won Kim
Department of Chemical Engineering, Konkuk University

I. 서론

컴퓨터가 공정에 도입되면서 방대한 양의 자료가 얻어지게 되었으며 이런 자료들은 서로 긴밀한 연관이 있을 수 있다. 또한 자료는 random error를 포함하며 측정기기의 파손 및 오동작 등에 의한 gross error를 포함할 수 있다. 이러한 자료를 분석함으로서 유용한 정보를 뽑아 내고 불필요한 정보를 제거해야 한다.

Gross error detection에 chi-square collective test (Reilly, 1963), univariate measurement test (Mah et al., 1982; Crowe et al., 1983), MP measurement test (Almasy, 1975), MP constraint test (Crowe, 1989; 1992) 등의 통계학적인 방법이 사용되고 있지만 만족할 만하게 gross error를 찾아내지 못하거나 정상적인 변수를 gross error를 포함하는 변수로 지적하는 경우가 있다.

최근 chemometric 방법중 Principal Component Analysis (PCA)와 Partial Least Squares 또는 Projection to Latent Structures (PLS) 방법이 multivariate analysis에 이용되고 있다. PCA의 개념은 Pearson (1901)에 의해서 도입되었고, Hotelling (1933)에 의해서 일반화되었으며 Jolliffe (1986), Wold (1987), Jackson (1991)등에 의해서 재확인되었다. Kresta (1991)는 연속 공정(continuous process)의 monitoring에 PCA/PLS방법을 도입했으며 Nomikos와 MacGregor (1994)에 의해 회분 공정(batch process)으로 확장되었다. MacGregor (1994), Tong과 Crowe (1995)에 의해 on-line data reconciliation에서 PC test로 발전됐다. PCA나 PLS 등의 분석 방법은 dimensionality reduction, measurement classification, outlier detection과 process monitoring, process modeling과 prediction 등에 이용되고 있다.

여러 논문에서 화공 분야의 제한적인 PCA의 적용에 대해서 언급했지만 data recolciliation에 관해서는 거의 적용된 것이 없었다. 본 논문에서는 PCA 또는 PLS 방법(chemometric methods)을 이용한 gross error detection과 공정 monitoring에 대해서 언급하고자 한다.

II. 이론

PCA의 주개념은 분산된 데이터를 몇 개의 축 상에 분포하도록 주요 축을 결정하는 데 있다. PCA 모델은 rank가 r인 X(known matrix 또는 predictor matrix)를 rank가 1인 r개 행렬(M)의 합으로 표현한다.

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + \ldots + \mathbf{M}_r$$

$$\mathbf{M}_h = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \ldots + \mathbf{t}_a \mathbf{p}_a^T$$

여기서 t는 score vector이며, 그 요소는 PC (Principal Component)선의 좌표를 나타낸다. p^T는 loading vector이며, 그 요소는 cosine방향의 요소, 즉 PC의 단위 벡터의 사영(projection)을 나타낸다.

이의 계산에는 X의 행렬의 크기가 작은 경우에는 Singular Value Decomposition

(SVD) algorithm이 사용되고 X의 행렬의 크기가 큰 경우에는 Nonlinear Iterative Partial Least Squares (NIPALS) algorithm이 사용되고 있다. 보다 일반적으로 널리 사용되는 반복법인 NIPALS algorithm은 아래와 같다.

① take a vector x_j from X and call it t : $t = x_j$
② calculate p^T : $p^T = t^TX / t^Tt$
③ normalize p^T : $p^T_{new} = p^T_{old} / \|p^T_{old}\|$
④ calculate t : $t = Xp / p^Tp$
⑤ compare the t used in step ② with that obtained in step ④.
If they are the same, stop. If they are still different, go to step ②.

Multiple Linear Regression (MLR)은 X-와 Y-Block (observed matrix 또는 predicted matrix)사이의 선형의 관계를 분석하는 것이며, Principal Component Regression (PCR)은 X-Block의 PCs와 Y-Block사이의 관계를 분석하는 것이다. PLS는 X-Block의 PCs와 Y-Block의 PCs사이의 관계를 분석하는 것이다. 이들의 관계 해석은 Linear PLS와 NonLinear PLS (Polynomial PLS, Spline PLS, Neural Net PLS 등)로 구분할 수 있다. NIPALS algorithm을 이용한 PLS algorithm은 아래와 같다.

$$X = TP^T + E = \Sigma t_h p_h^T + E$$
$$Y = UQ^T + F = \Sigma u_h q_h^T + F$$
$$u_h = b_h t_h \rightarrow b_h = u_h^T t_h / t_h^T t_h \text{ (linear relation)}$$

① take a vector y_h from Y and call it u : $u = y_h$
② calculate w^T : $w^T = u^TX / u^Tu$
③ normalize w^T : $w^T_{new} = w^T_{old} / \|w^T_{old}\|$
④ calculate t : $t = Xw / w^Tw$
⑤ calcuate q^T : $q^T = t^TY / t^Tt$
⑥ calcuate u : $u = Yq / q^Tq$
⑦ compare the u used step ② with that obtained in step ⑥.
If they are the same, go to step ⑧. If they are still different, go to step ②.
⑧ calculate p^T : $p^T = t^TX / t^Tt$
⑨ calculate inner relation : linear, nonlinear, spline, neuralnet ...
⑩ calcuate residuals : $E = X - tp^T$, $F = Y - tq^T$

여기서 t와 u는 score vector를 p^T와 q^T는 loading vector를 w^T는 weight vector를 E와 F는 residual matrix를 나타낸다. b는 linear relation coefficient이다.

III. 결과 및 토론

Chemometric 방법 (PCA/PLS)을 이용한 gross error detection을 위해서 data generation 과 preprocessing, normal operating condition, gross error detection and identification의 과정으로 검토했다.

Data generation 과 Preprocessing: Phillips(1991)의 berty reactor모델의 측정 데이터 중 data reconciliation과 gross error detection된 data set에서부터 수지식 (balance

Latent Variables	PCA(%)			PLS(%)	
	X	Y	X+Y	Linear	Spline(nonlinear)
1	72.7	62.9	69.3	X:72.1 Y:41.8	X:72.1 Y:42.4
2	84.7	78.7	83.1	X:84.5 Y:69.1	X:84.4 Y:71.4
3	93.2	87.2	91.1	X:93.1 Y:71.7	X:93.0 Y:74.1
4	97.6	92.4	95.5	X:97.5 Y:75.3	X:97.6 Y:77.1
5	98.7	94.9	97.2	X:98.7 Y:80.4	X:98.6 Y:83.6

Table 1. Cumuative % Sum of Squares Explained by the PCA and PLS

equation) 과 반응식을 이용해 모사 데이터를 만들었다. PCA/PLS의 적용을 위해서 모사 데이터에 random error를 더했으며, mean centering 처리했다. 이 데이터를 normal operating condition (NOC)로 가정해서 PCA/PLS 과정을 수행했다. Gross error가 검사되는지 확인하기 위해서 임의의 data set(33)의 4번째 변수에 gross error에 해당하는 큰 값을 주었다.

PCA/PLS: 경우에 따른 gross error detection 정도를 알아보기 위해서 PCA의 경우 X, Y, XY-block으로 나누어 검토했으며 PLS의 경우 모델의 비선형성때문에 linear-PLS를 적용했을 때보다 nonlinear-PLS를 적용했을 때 보다 더 효과적이었다. 각각의 경우에 대한 결과는 Table. 1에 표시하였다.

Detection 과 Identification: Gross error detection을 위한 NOC의 montoring chart (Fig. 1 and Fig. 2)를 그렸다. Fig. 1은 2개의 Latent Variables (LV)에 대한 score chart이며, 점선과 실선은 각각 95%와 99% confidence region을 표시한다. Fig. 2는 각 변수에 대한 Squared Prediction Error (SPE) chart이다. 측정데이터에 gross error가 존재할 경우 score chart나 SPE chart에 나타나게 된다. Fig. 1과 Fig. 2에서 33번째 data set에 gross error가 포함되었음을 나타내고 있다. Monitoring chart를 통해서 gross error가 포함된 데이터가 검출됨을 알 수 있다. SPEx 값에 영향을 준 변수를 구분하기 위해서 33번째 data set의 Prediction Error (PE) chart (Fig. 3)를 도입했으며 이 chart로부터 1, 4, 5번째 변수에 문제가 있음을 알 수 있다. 하지만 정확히 어느 변수 또는 몇 개의 변수에 문제가 있는지 확인할 수 없다. 보조적으로 문제가 있는 변수를 알아보기 위해서 33번째 data set의 Identification (ID) chart (Fig. 4)를 도입했다. ID chart는 PCA/ PLS 방법에서 첫 LV에 많은 값이 잡히기 때문에 $LV_1(t_1)$방향으로의 각 공정 변수의 변화량을 표시하였다. Fig. 4에서 보듯이 4번 변수의 영향으로 33번째 data set이 t_1방향으로 이동했음을 알 수 있다.

여러 경우의 gross error에 대해 실험한 결과 Philips가 제시한 방법보다 detection 과 identification 효율이 좋은 것으로 나타났다. 공정에서 얻어진 raw data를 PCA/PLS를 이용해 gross error detection 과정을 수행한 후 검출된 gross error를 바탕으로 Data reconciliation하는 새로운 algorithm이 제시될 수 있겠다.

IV. 참고문헌

Geladi, P., and B. R. Kowalski, "Partial Least-Squares Regression: A Tutorial," *Analytica Chimica Acta*, **185**, 1 (1986).

Kaspar, M. H., and W. H. Ray, "Chemometric Methods for Process Monitoring and High-Performance Controller Design," *AIChE J.*, **38**, 1593 (1992).

Kresta, J. V., J. F. MacGregor and T. E. Marlin, "Multivariate Statistical

Monitoring of Process Operating Performance," *Can. J. Chem. Eng.*, **69**, 35 (1991).

MacGregor, J. F., C. Jaeckle, C. Kiparissides and M. Koutoudi, "Process Monitoring and Diagnosis by Multiblock PLS Methods," *AIChE J.*, **40**, 826 (1994).

Mardia, K., J. Kent and J. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.

Nomikos, P., and J. F. MacGregor, "Monitoring Batch Processes using Multiway Principal Component Analysis," *AIChE J.*, **40**, 1361 (1994).

Phillips, A. G., *Application of Data Reconciliation and Gross Error Detection to a Reaction Rate Modeling Problem*, Master's thesis, West Verginia Univ., USA, 1991.

Tong, H., and C. M. Crowe, "Detection of Gross Errors in Data Reconciliation by Principal Component Analysis," *AIChE J.*, **41**, 1712 (1995).

Wold, S., "Discussion: PLS in Chemical Practice," *Technometrics*, **35**, 136 (1993).

Wold, S., K. Esbensen and P. Geladi, "Principal Component Analysis," *Chemometrics Intell. Lab. Syst.*, **2**, 37 (1987).

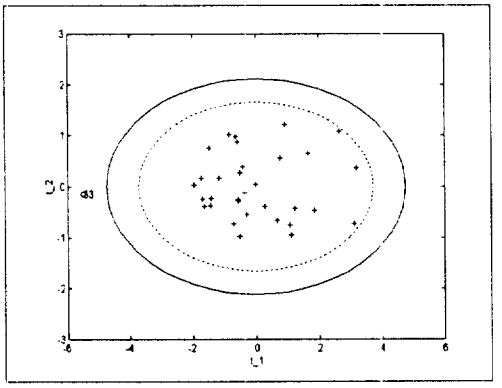

Fig. 1. Score Chart; t_1-t_2 plane; A gross error exists in data set 33

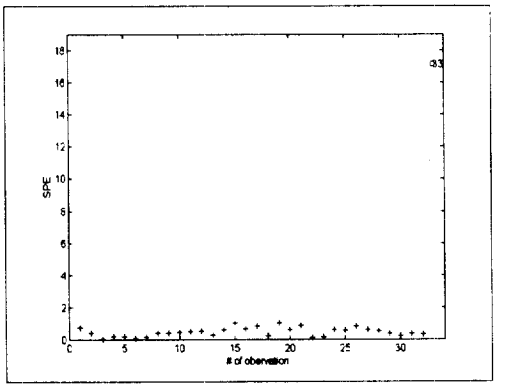

Fig. 2. SPEx Chart; A gross error exists in data set 33

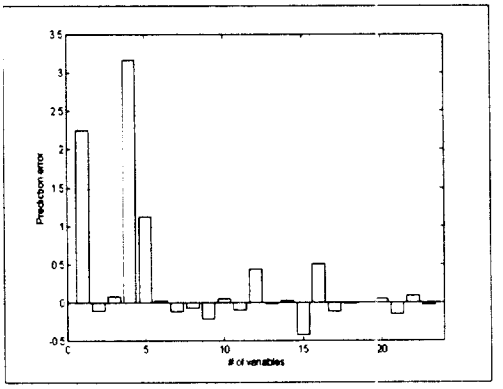

Fig. 3. PE Chart; Pridiction errors in the individual process variables contributing to SPEx at data set 33

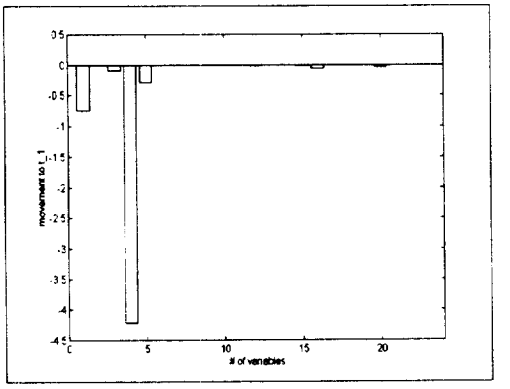

Fig. 4. ID Chart; Variable contributions to the change in t_1 from center of t_1-t_2 plane