동특성과 다변량 통계기법

지난 20 년간 DCS 및 센서 기술의 발달과 보급으로 인해 오늘날엔 보다 많은 양의, 그리고 많은 상관관계(correlations)를 가진 데이터가 공장의 데이터 베이스로 모아진다. 이는 기존의 한 두가지 품질변수만을 모니터링하던 방법들(SPC 등)에게 해결하기 힘든 문제점들을 부과하였으며 따라서 데이터를 저차원의 공간에 투영하여 데이터의 크기와 상관관계를 줄이는 다변량 통계기법이 각광을 받기 시작하였고 지금까지 많은 산업현장에 성공적으로 적용되었다.

Wise 등[4]이 다변량 통계기법의 시스템 동특성 적용에 대한 이론적 기초를 제공하였으나, 그 이후에도 이(다변량 통계기법의 시스템 동특성 적용)에 대한 많은 문제점과 연구가 제기되었다. 이 글에서는 기존 문헌들의 정리, 비교를 통해 이 문제점에 대하여 다시 생각해 보는 기회를 마련하고자 한다.

1. 개선된 다변량 통계기법

기존의 다변량 통계기법들이 시스템의 동특성을 포함한 많은 경우에 성공적으로 적용되었음에도 불구하고 여러 방법들이 이를 개선하여 시스템의 동특성을 표현하고자 노력하였다. 그 중 대표적인 것이 동적(dynamic) PCA 와 canonical variate analysis (CVA) 다.

(1) 동적 PCA (DPCA) [1]

동적 시스템의 경우, 변수의 현재 값은 과거 값에 의존한다. 따라서 최소한 $\mathbf{X}(k)$ 과 $\mathbf{X}(k\text{-}1)$ 사이의 선형 동적관계를 확인해야 한다. 만약 모든 동적 관계가 1 차인 시스템이면, 그 관계는 다음의 잡음(noise) 부공간(subspace) 이다.

$$[\mathbf{X}(k)\ \mathbf{X}(k-1)]\mathbf{b} = 0 \qquad (1)$$

이때 $\mathbf{X}$ 는 $m\times n$ 행렬이고 $\mathbf{b}$ 는 우측-singular 벡터이다. 더 일반적인 경우엔,

$$\mathbf{X}_A(l)\mathbf{b} = 0 \qquad (2a)$$

이고 이때

$$\mathbf{X}_A(l) = [\mathbf{X}(k)\ \mathbf{X}(k-1) \cdots \mathbf{X}(k-l)] \qquad (2b)$$

이다. 행렬 $\mathbf{X}_A(l)$는 모델 차수감소와 시스템 인식에 널리 쓰이는 Hankel 행렬의 특수한 형태를 취하고 있다. 이것은 데이터 행렬이 시간 전이된 벡터들로 이루어 진것을 제외하고는 궁극적으로 기존의 PCA 방법과 동일하다. 적절히 l 값을 선택함으로서, 정적 및 동적 관계 모두 작은 singular 값들을 가지고 잡음공간에 나타나야 한다. 잡음공간으로부터 계산된 *Q statistics* 는 독립적이며 동적 시스템에 대한 통계적 기반을 제공한다.

Ku 등[1]은 DPCA 에서 가장 중요한 l 값을 결정하는 몇 가지 방법을 제시하였지만, 제시된 방법들은 데이터에 잡음이 있을 경우 정확한 l 값을 결정하지 못하고, 변수의 개수가 많을 때는 수행하기가 힘들며, 또한 그 결과를 해석하기가 어렵다는 단점이 있다.

(2) Canonical Variate Analysis (CVA) [2]

CVA 는 suboptimal 인식(identification) 알고리즘의 한 종류이다. CVA 의 idea 는 latent 변수의 개수를 정하는데 있어 covariance 와 cross-covariance 정보를 사용한다는 점에서 PCA, PLS 와 유사하다. PLS 는 cross-covariance 를 최대화하고 순차적으로 중요한 latent 변수를 결정하는 반면, CVA 는 cross-correlation 을 최대화하며 SVD (Singular Value Decomposition)을 사용한다. Canonical correlation analysis 는 통계학 분야에서 시계열 모델링에 사용되어 왔다. 이것은 필터링된 과거 출력으로부터 현재의 공정출력의 예측을 최대화한다. 그러나 CVA 는 과거

공정출력과 미래의 공정출력 사이의 canonical correlations 을 최대화한다. CVA 를 통한 상태-공간 모델링은 다음과 같다:

Hankel 행렬 $\mathbf{H}_{JK}$, 상태 z_k, 무작위오차 e_k 를 다음과 같이 정의한다:

$$H_{IJ} = E(y_{K_J}^{+} y_{K-1_J}^{-T}) = E\left\{\begin{bmatrix} y_k \\ \vdots \\ y_{K+J-1} \end{bmatrix}\begin{bmatrix} y_{k-1}^T & \cdots & y_{k-K}^T \end{bmatrix}\right\} \tag{3}$$

$$z_k = E\left(x_k \middle| y_{k-1_K}^{-}\right),\ e_k = y_k - E\left(y_k \middle| y_{k-1_K}^{-}\right) \tag{4}$$

그러면 상태벡터의 covariance 행렬은 미래와 과거의 canonical variates 의 covariance 행렬과 같다.

$$E(z_k z_k^T) = \Sigma,\ \overline{\mathbf{H}}_{JK} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \tag{5}$$

이때 $\overline{\mathbf{H}}_{JK}$ 는 scaled Hankel 행렬이다. 상태-공간 모델에서 시스템 행렬 (**A**, **B** 와 **C**)는 최소 자승법으로 구할 수 있다.

엄격히 말해서, 이 방법은 다변량 통계기법이라기 보다는 다변량 시스템 인식 기법이라고 보는 것이 옳다고 생각되며, 일단 상태-공간 모델을 구하고 나면 선형시스템 이론의 많은 결과들을 사용할 수 있다는 장점이 있다. 주된 단점 중 하나는 기존의 다변량 통계기법보다 잘 성계된 시스템 인식 실험이 필요하다는 것이다.

2. 동적 공정의 PCA 모니터링 [3]

PCA 와 상태-공간 모델 출력식의 유사성을 이용하여 wise 등은 잡음은 주로 (PCA 모델의) 나머지 (residuals)에 영향을 미치는 반면, 공정 상태의 변화는 주로 PCA score 의 변화로 나타난다는 것을 보였다. 이것은 PCA 나머지에 대한 오차 판별 상한선을 도출해 낼때 사용자로 하여금 잡음 성질만을 관찰할 수 있게 한다.

다음과 같은 선형 시불변 이산 상태 모델로 표현되는 '즉시 관측 가능한' 공정을 고려해 보자. ('즉시 관측 가능한' (*immediately observable*) 상태란 p 개의 출력 중 하나의 표본으로부터 추정이 가능한 n 개의 상태 변수의 부분 집합으로 정의된다.)

$$\begin{aligned} x(k+1) &= \mathbf{A}x(k) + \mathbf{B}u(k) + e(k) \\ y(k) &= \mathbf{C}x(k) + e(k) \end{aligned} \tag{6}$$

그러면 출력 식은 다음과 같고

$$y(k) = \mathbf{P}_a x_q(k) + e(k) \tag{7}$$

이때 $\mathbf{P}_q$ 는 orthonormal 행렬이고 q 는 즉시 관측 가능한 상태 개수이다. 주어진 출력 표본 $y(k)$ 와 식 (7)을 가지고 다음과 같이 측정 잡음을 추정할 수 있다.

$$\begin{aligned} \hat{e}(k) &= y(k) - \hat{y}(k) \\ &= (\mathbf{I} - \mathbf{P}_q \mathbf{P}_q^T) y(k) \end{aligned} \tag{8}$$

이때

$$\begin{aligned} y(k) &= \mathbf{P}_q x_q(k) \\ &= \mathbf{P}_q \mathbf{P}_q^T y(k) \end{aligned} \tag{9}$$

만약 m 개의측정치가 있으면, 측정된 출력 $\mathbf{Y}(m \times p)$ 과 상태 $\mathbf{X}(m \times n)$, 잡음 $\mathbf{E}(m \times p)$ 사이의 관계는

$$\mathbf{Y} = \mathbf{X}\mathbf{C}^T + \mathbf{E} \tag{10}$$

이다. q 개의 latent 변수를 갖는 $\mathbf{Y}$ 에 대한 PCA 분해는 다음과 같다.

$$\mathbf{Y} = \mathbf{T}_q \mathbf{P}_q^T + \mathbf{E}_{p-q} \tag{11}$$

식(6)과 (10)을 비교해 보면, PCA 는 q 개의 즉시 관측 가능한 상태 $\mathbf{T}q$ 와 행렬 $\mathbf{C}$ 를 대신하는 $\mathbf{P}q$ 로 이루어진 상태-공간 모델로 해석될 수 있다. 주어진 출력 표본 $y^{\mathrm{T}}(k)$ 에 대한 PCA 나머지는 다음과 같이 정의된다.

$$r(k) = (\mathbf{I} - \mathbf{P}_q \mathbf{P}_q^T) y(k) \tag{12}$$

따라서, $\mathbf{P}_q$ 가 정확하다면 PCA 나머지는 식(8) 에서 정의된 추정된 잡음과 동일하다. 다시 말해서, 나머지는 시스템의 동특성과 상관없이 $e(k)$ 만의 함수이다.

PCA 모니터링의 실패와 성공은 모델(이 경우 $\mathbf{P}_q$ 행렬)의 정확도에 달려있다. 따라서, 정확한 PCA 모델을 위해선 잘 설계된 시스템 인식 실험이 필요하다. 그렇지 않을 경우, Autocorrelation 을 갖는 데이터는 무작위 잡음이라고 간주하고 계산하는 Q statistics 의 제어극한 계산에 영향을 준다.

4. 토론

PCA 와 PLS 등 기존의 다변량 통계기법들이 동적 시스템에도 성공적으로 적용된 사례가 있음에도 불구하고, 위에 설명된 것과 같은 개선된 방법들이 제안된 이유 중 하나는 기존의 방법을 장기간 사용했을 때 잦은 false alarm 을 유발한다는 점이다. 왜냐하면, 화학공정의 특성상 갖는 여러가지 다른 조업 조건을 하나의 PCA/PLS 모델에 반영한다는 것은 현실적으로 매우 힘들며, 계절에 따른 냉각수 및 외부 온도 같은 외부 조건의 변화 또한 반영하기 힘들기 때문이다. 이같은 문제점은 DPCA, CVA 같은 개선된 방법에도 그대로 해당된다. DPCA 는 결국 단변량(univariate) 또는 다변수(multivariable) 시계열 모델링을 다변량의 경우로 확장한 것에 지나지 않으며, 위에서 언급한 바와 같이 CVA 는 다변량 시스템 인식 기법의 하나이기 때문에 두 방법 모두 단기간 적용시에는 기존의 방법보다 좋은 성능을 보일 수도 있지만 (Russel 등[4]의 벤치마킹 결과에 따르면 PCA 와 DPCA 는 거의 동일한 성능을 보임.) 이들 역시 위와 같은 공정변화에는 마찬가지 결과를 보인다 (DPCA 의 l 의 경우 생산량(유량)에 따라 변함.). 이를 극복하기 위해(예를 들면) 냉각수의 온도 대신 냉각수와 반응기의 온도차 등을 이용하기도 하지만 궁극적인 해결책은 PCA/PLS 모델을 매 샘플링때마다 갱신하는 adaptive PCA/PLS 를 사용하거나[5] wavelet 분해를 이용, 데이터에서 계절성향 등 장기성향을 제거한 뒤 PCA/PLS 를 적용하는 것[6]이 바람직하다고 생각된다.

참고문헌

1. Ku, W., Storer, R.H. and Georgakis, C., Chemometrics and Intelligent Laboratory Systems, 30 179(1995).
2. Negiz, A. and Cinar, A., AICHE J, 43(8), 2002(1997).
3. Wise, B.M., Ricker, N.L., Veltkamp, D.F. and Kowalski, B.R., Process control and quality, 1, 41(1990).
4. Russel, E.L., Chiang, L.H. and Braatz R.D., Chemometrics and Intelligent Laboratory Systems, 51 81(2000).
5. Li, W., Yue, H., Sergio, V.C. and Qin, S.J., Journal of Process Control, 10 471(2000).
6. Teppola, P. and Minkkinen, P., Journal of Chemometrics, 14 383(2000).